Efficient reinforcement learning with Fleming-Viot particle systems: application to stochastic networks with rarely observed rewards

Daniel Mastropietro, Urtzi Ayesta, Matthieu Jonckheere, Szymon Majewski

Efficient reinforcement learning with Fleming-Viot particle systems: application to stochastic networks with rarely observed rewards

Daniel Mastropietro[1], Urtzi Ayesta[2], Matthieu Jonckheere[3], and Szymon Majewski[4]

[1]CNRS-IRIT, Université de Toulouse INP, 31071 Toulouse, France
[2]CNRS-IRIT, 31071 Toulouse, France / Ikerbasque - Basque Foundation for Science, 48009 Bilbao, Spain / UPV/EHU, University of the Basque Country, 20018 Donostia, Spain
[3]CNRS-LAAS, CNRS-IRIT, 31071 Toulouse, France
[4]Ecole Polytechnique, Paris, France

July 14, 2023

Abstract

We consider reinforcement learning control problems under the expected reward criterion in which non-zero rewards are both sparse and rare, that is, they occur in very few states and have a very small stationary probability under all policies. In this context, usual discovery techniques including importance sampling are inapplicable because no policy exists that increases the visit frequency of the rare states. Using renewal theory and Fleming-Viot particle systems, we propose a novel approach that exploits prior knowledge on the sparse structure of the reward landscape to boost exploration of the rare non-zero rewards and achieve an accurate estimation of their stationary probability. We also demonstrate how to combine the methodology with policy gradient learning to construct the FVRL algorithm that efficiently solves control problems under these scenarios. We provide theoretical guarantees of the convergence of both the stationary probability estimator and the policy gradient learner, and illustrate the method on two optimisation problems to maximize the expected reward: a simple $M/M/1/K$ queue system where the blocking threshold K is optimised, and a two-job-class loss network where a threshold-type rejection policy is optimised. Our results show that FVRL learns the optimum thresholds much more efficiently than vanilla Monte-Carlo reinforcement learning.

Keywords: semi-Markov decision process, parameterised policy, policy gradient, queues, loss network

1 Introduction

Reinforcement learning methods, by being able to take advantage of large amounts of computational resources, have been very successful at solving very complex problems with large state dimensions and sparse rewards, obtaining super-human performance particularly in games [23]. Many of these successes have been obtained in an episodic setting in which, even under sparse rewards conditions, there is certainty that the episode will eventually finish, at which moment a reward will be observed.

In several application domains, in particular in networking or robotics, the environment is not episodic (i.e. there is no notion of progression as in games) and non-zero rewards are received only in a handful of states which are visited very rarely. We refer to these as environments with sparse and rare rewards. For example in networking, a fundamental problem is how to dimension the system in order to optimise the performance, bearing in mind that the blocking probability (i.e. the probability that the system cannot accept a new data packet, call, or computation task) can be extremely small. As pointed out in [11], efficiently managing the exploration task when non-zero rewards are very rarely observed remains a challenge, and this provides the main motivation for the present work. Our starting point is the structural knowledge on the underlying Markov Decision Process (MDP) with which the system is modelled, that in many cases can be leveraged to drastically improve exploration.

In this paper, we focus on model-free approaches (i.e. without previous learning of a specific model), and take as performance criterion the expected reward (or expected cost, when more natural for the problem at hand). We assume that we have access to a simulator or emulator of the system, or in its absence, to a large amount of data to construct experience replay of the system. Our approach relies on the identification of sets $\mathcal{A}$ of states with zero reward, which can be obtained through previous knowledge on the underlying MDP, or from information gained from exploration already performed.

We first show that the expected reward of the original problem can be expressed in terms of modified trajectories that are absorbed in $\mathcal{A}$. We then show that the expected reward of trajectories outside $\mathcal{A}$ can be efficiently estimated via the so-called Fleming-Viot particle system (FV). We finally introduce FVRL, a reinforcement learning algorithm that uses Fleming-

Viot to solve control problems with sparse and rare rewards and gradients.

To illustrate the main ideas, we shall consider an $M/M/1/K$ queue, where the objective is to minimize a convex expected cost on blocking events. This is a fundamental model in queueing theory and an interesting archetype of dynamics where the rewards (here seen as costs) can only be observed during unlikely excursions for system loads smaller than 1. More concretely, the stationary probability of being in state K decreases exponentially with K at rate ρ^K, where ρ is the load of the system, assumed smaller than 1. The rewards are thus sparse, since blocking only occurs when the system is full, and rare since this happens with very low probability for a wide range of values of K and ρ that are common in practice.

Using this example, the intuition behind our proposal relies on the fact that the probability of visiting state K conditioned on not having visited state 0, which decays as $\sqrt{K}\rho^{K/2}$ when K increases [10], is considerably larger than the stationary probability of visiting K. For example, for $K = 40$ and $\rho = 0.7$, the conditioned probability is about 8 thousand times larger than the unconditioned one, and about 6 million times larger when $\rho = 0.5$. Hence, by shifting the estimation toward a conditioned dynamics, which we propose to achieve using Fleming-Viot particle systems, we expect a lower sample complexity, both for estimation and for control. Our numerical results show that the Fleming-Viot approach overcomes the difficulty of vanilla Monte-Carlo to properly estimate key quantities for reinforcement learning algorithms, such as value functions and gradients.

In summary, we propose a methodology that can be used to boost reinforcement learning in environments with the following characteristics: (i) they present sparse non-zero rewards and sparse gradients of parameterised policies occurring very rarely at unknown states; (ii) thanks to prior knowledge, it is possible to define a set of states presenting zero rewards. The learning goal considered here is to maximize the expected reward under stationary regime operation but this could be easily generalised to contexts with discounts. In environments where the occurrence of non-zero rewards has an extremely low probability, this methodology enables learning at admissible learning times, whereas vanilla Monte-Carlo methods fail.

In our main contribution, we develop FVRL, a method that combines FV and RL in problems with sparse and rare rewards that are ubiquitous in the control of stochastic networks. In our main theoretical contributions for FVRL, we establish convergence of the estimator as the number of particles tends to infinity, and convergence to the optimal policy. Our numerical results on a simple queueing model and a multi-class loss network illustrate the potential benefits of the approach, and we will investigate in further research its applicability to a wider variety of examples, including classical RL environments.

We would also like to note that, to the best of our knowledge, there is no adequate solution to this problem in the state of the art. Even though we do not focus here on deep learning tools, our proposal could definitely be used in combination with them. This is left for future work. Initially, one might think that the problem of estimating rare events could be tackled using Importance Sampling (IS), an area in which there exists a large literature. However, it is important to observe that the problem at hand impedes the application of IS methods. Firstly, we here look at scenarios where **there is no policy** allowing to explore rare states. In other words, states with non-zero rewards are very rarely explored under all policies, so IS is not an option, as it is designed to change the exploration policy. Secondly, it is also not possible to invoke the original IS principle when looking purely at the evaluation task for a fixed policy, as this requires a change of measure for Monte-Carlo, which is not possible in our case as the transition rates of the Markov process are unknown.

On the other hand, the problem of sparse rewards has been of central interest for a long time and the most intuitive solution to sparse (but not necessarily rare) reward problems is reward shaping. Mataric formulated the idea back in 1994 [17]. However, these methods have a few drawbacks: expertise is needed to shape the rewards, and only very few and quite arbitrary policies might be reached as a consequence of shaping. Our proposal fits more into the idea of curiosity-driven methods as explained for instance in [18]. The idea is basically to encourage the exploration of unvisited states in the environment. While existing mechanisms are based on including a bonus term in the loss function to favour exploration ([18, 6]), we define a new (more radical) curiosity mechanism adapted to sparse and rare reward environments that forces exploration outside a set of states that have already been explored or are known to be uninformative.

The rest of the paper is organized as follows. Section 2 describes the mathematical setting of the problem, Section 3 the general methodology, in Section 4 we show its applicability in the cases of an $M/M/1/K$ queue system and of a loss network serving two classes of jobs.

2 Problem description

We consider a continuous-time MDP $(\mathcal{S}, A, q, \mathcal{R})$ with a finite state space $\mathcal{S}$, action space A, jump rates q, and rewards $\mathcal{R}$, under the expected reward criterion. Throughout the paper we assume that for each policy π the continuous-time Markov process X_t^π obtained by following the policy π is irreducible. We denote by p^π the stationary distribution of X_t^π, and by $\mathbb{E}^\pi(\eta)$ the expectation with respect to p^π of a function of interest, $\eta : \mathcal{S} \to \mathbb{R}$, such as the occupation measure or the reward.

We will be interested in computing $\mathbb{E}^\pi(\eta)$ under the assumption that the function η is zero outside of a small set of states $\mathcal{C} \subset \mathcal{S}$. For example, if the rewards are assumed to be sparse, the reward function $r : \mathcal{S} \times A \to \mathbb{R}$ is zero outside a small set $\mathcal{C} \times A$, from which we can define the function η as $\eta(x) = \sum_a r(x,a)\pi(a|x)$, whose expected value can be reduced to a calculation over the states in set $\mathcal{C}$ as $\mathbb{E}^\pi(\eta) = \sum_{x\in\mathcal{C}} p^\pi(x)\eta(x)$.

The objective of computing $\mathbb{E}^\pi(\eta)$ will be achieved as follows: we will first choose a set $\mathcal{A} \subset \mathcal{S}$, such that $\mathcal{C} \cap \mathcal{A} = \emptyset$. Then, we will use Fleming-Viot particle systems to compute the expectation truncated to $\mathcal{A}^c$, $\sum_{x\in\mathcal{A}^c} p^\pi(x)\eta(x)$, which is equal to $\mathbb{E}^\pi(\eta)$ given that η is zero in $\mathcal{A}$.

3 Methodology

In this section we present a new method to estimate $\mathbb{E}^\pi(\eta)$ tailored to Markov decision processes with sparse and rare rewards. We then discuss how to use this algorithm to improve the estimation of gradients in the context of the policy gradient methodology to solve optimal control problems with sparse and rare rewards. For simplicity of exposition, we assume that we have direct access to the function η. We also assume throughout this section that the policy π and the chosen set $\mathcal{A} \subset \mathcal{S}$, whose intersection with $\mathcal{C}$ is empty, are fixed.

3.1 Definitions

The following definitions will be instrumental in our discussion. We denote by $\vec{\partial}\mathcal{A}$ the entrance boundary of $\mathcal{A}$, i.e. the set of states $x \in \mathcal{A}$ for which there exists at least a state $y \in \mathcal{A}^c$ with positive jump rate to x, i.e. $q(y,x) > 0$. The entrance boundary of $\mathcal{A}^c$, $\vec{\partial}\mathcal{A}^c$, is defined analogously. We define the first time of entry into $\mathcal{A}$ as

$$\mathcal{T}_{\mathcal{A},0} \doteq \inf\{t > 0 : X_t^\pi \in \mathcal{A} \text{ and } X_{t^-}^\pi \notin \mathcal{A}\},$$

and we denote the entrance state distribution into $\mathcal{A}$ under stationarity as:

$$p^\pi_{\vec{\partial}\mathcal{A}}(x) \doteq \mathbb{P}(X^\pi_{\mathcal{T}_{\mathcal{A},0}} = x | X_0^\pi \sim p^\pi), \forall x \in \vec{\partial}\mathcal{A}.$$

Using Figure 1 as a visual aid, we further define the first time of entry into $\mathcal{A}^c$ following $\mathcal{T}_{\mathcal{A},0}$ as $T_{\mathcal{A}^c} \doteq \inf\{t > \mathcal{T}_{\mathcal{A},0} : X_t^\pi \in \mathcal{A}^c\}$, the first time of entry into $\mathcal{A}$ following $T_{\mathcal{A}^c}$ as $T_{\mathcal{A}} \doteq \inf\{t > T_{\mathcal{A}^c} : X_t^\pi \in \mathcal{A}\}$, and their difference as $T_{\mathcal{K}} \doteq T_{\mathcal{A}} - T_{\mathcal{A}^c}$, also referred to as the killing time. The stopping time $T_{\mathcal{A}}$ will be regarded in the sequel as a cycle return time to $\mathcal{A}$.

Finally, for any measurable subset B, we define $\mathbb{P}_{\vec{\partial}\mathcal{A}}(B) \doteq \mathbb{P}(B | X_0^\pi \sim p^\pi_{\vec{\partial}\mathcal{A}})$. For the complement set $\mathcal{A}^c$, we define the entrance state distribution into $\mathcal{A}^c$ under stationarity as:

$$p^\pi_{\vec{\partial}\mathcal{A}^c}(x) \doteq \mathbb{P}_{\vec{\partial}\mathcal{A}}(X^\pi_{T_{\mathcal{A}^c}} = x), \forall x \in \vec{\partial}\mathcal{A}^c$$

.

The two state probability distributions defined above, $p^\pi_{\vec{\partial}\mathcal{A}}(x)$ and $p^\pi_{\vec{\partial}\mathcal{A}^c}(x)$, will be thoroughly used in the development of our proposed methodology to condition the start state of the Markov decision process X_t^π.

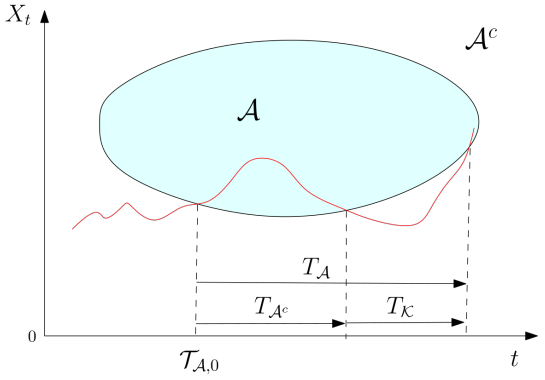


Figure 1: Stopping times defined in Section 3.1.

3.2 Estimation with Fleming-Viot particle systems

We now introduce the Fleming-Viot method as a way to tackle the problem of sparse and rare rewards and estimate $\mathbb{E}^\pi(\eta)$ efficiently. Our approach penalizes trajectories that enter $\mathcal{A}$ in order to boost exploration of the subset $\mathcal{C}$ of the state space that is relevant for the estimation of the expectation, which is in $\mathcal{A}^c$. The penalisation consists in immediately replacing trajectories that enter $\mathcal{A}$ by trajectories outside $\mathcal{A}$. To this end, we use the dynamics of a particle system known as Fleming-Viot (FV) [7] which has been used in the literature to simulate quasi-stationary distributions [7, 2].

More specifically, the Fleming-Viot N-particle system with driving process X_t^π and absorption set $\mathcal{A}$ is a continuous-time Markov process $(\xi_t^\nu)_{t\geq 0}$ defined on the state space $(\mathcal{A}^c)^N$ as follows: given a probability distribution ν on $\mathcal{A}^c$, an N-dimensional vector $\xi_0^\nu(k)_{k=1,\dots N}$ defines the initial state of the particles in the system and is obtained as i.i.d. samples from ν. Each particle $\xi_t^\nu(k)$ then evolves independently according to the dynamics of X_t^π, but whenever it hits a state in $\mathcal{A}$, it immediately jumps to the position of one of the other particles chosen uniformly at random. This mechanism allows us to only explore trajectories outside $\mathcal{A}$, which is where the informative rewards are located.

In order to exploit the Fleming-Viot particle system for the estimation of $\mathbb{E}^\pi(\eta)$, we leverage the renewal theory characterization of the stationary probability of a state in terms of return cycles [1, chapter 6, Theorem 1.2], as follows: if we use the entrance to $\mathcal{A}$ as the event defining the beginning and end of a cycle, the stationary expectation of an arbitrary function η can be written in terms of the cycle time $T_{\mathcal{A}}$ as

$$\mathbb{E}^\pi(\eta) = \frac{\mathbb{E}_{\vec{\partial}\mathcal{A}}\left(\int_0^{T_{\mathcal{A}}} \eta(X_t^\pi) dt\right)}{\mathbb{E}_{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}}. \tag{1}$$

Note that, if we used Monte-Carlo to estimate this expectation by simulating the Markov process X_t^π starting at $\vec{\partial}\mathcal{A}$ and observing cycle times $T_{\mathcal{A}}$, it might take a very long time before observing a non-zero contribution from η, as this function is assumed to be sparse with non-zero values rarely observed. The following proposition is key to define the Fleming-Viot estimation method.

Proposition 1. Given a set $\mathcal{A} \subset \mathcal{S}$ and a function $\eta : \mathcal{S} \to \mathbb{R}$ that is zero on $\mathcal{A}$, the following holds:

$$\mathbb{E}^\pi(\eta) = \frac{\int_0^\infty \mathbb{E}_{\vec{\partial}\mathcal{A}^c}\left(\eta(X_t^\pi)\mathbf{1}_{T_{\mathcal{K}}>t}\right) dt}{\mathbb{E}_{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}}, \tag{2}$$

which, for estimation purposes, can be further simplified as:

$$\mathbb{E}^\pi(\eta) = \int_0^\infty f_\eta(t) g(t) dt, \tag{3}$$

where

$$f_\eta(t) \doteq \sum_{x \in \mathcal{A}^c} \eta(x) \phi_t^{\vec{\partial}\mathcal{A}^c}(x); g(t) \doteq \frac{\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)}{\mathbb{E}_{\vec{\partial}\mathcal{A}} T_{\mathcal{A}}},$$

and $\phi_t^{\vec{\partial}\mathcal{A}^c}(x) \doteq \mathbb{P}_{\vec{\partial}\mathcal{A}^c}(X_t^\pi = x | T_{\mathcal{K}} > t)$ is the probability that the process X_t^π, started at a state in $\vec{\partial}\mathcal{A}^c$ chosen with probability $p^\pi_{\vec{\partial}\mathcal{A}^c}$, is in x provided it has not been absorbed.

Proof. Since at the beginning of a cycle the process starts at a state in $\vec{\partial}\mathcal{A}$, in order to reach a state in the set of interest $\mathcal{C} \in \mathcal{A}^c$ (the set of states with non-zero values of η), the process needs to go through $\vec{\partial}\mathcal{A}^c$.

Thus, using the definition of the sojourn times $T_{\mathcal{A}^c}$ and $T_{\mathcal{A}}$ and recalling that $T_{\mathcal{A}} > T_{\mathcal{A}^c}$, expression (1) can be written as

$$\mathbb{E}^\pi(\eta) = \frac{\mathbb{E}_{\vec{\partial}\mathcal{A}}\big(\int_0^{T_{\mathcal{A}^c}} \eta(X_t^\pi)dt + \int_{T_{\mathcal{A}^c}}^{\infty} \eta(X_t^\pi)\mathbf{1}_{T_{\mathcal{A}}>t}dt\big)}{\mathbb{E}_{\vec{\partial}_{\mathcal{A}}} T_{\mathcal{A}}}.$$

Since by assumption η is zero in $\mathcal{A}$, the first integral is zero. After changing the integration variable to $u = t - T_{\mathcal{A}^c}$ in the second integral and using the notation of the killing time $T_{\mathcal{K}} = T_{\mathcal{A}} - T_{\mathcal{A}^c}$, we get

$$\mathbb{E}^\pi(\eta) = \frac{\mathbb{E}_{\vec{\partial}\mathcal{A}}\big(\int_0^{\infty} \eta(X^\pi_{u+T_{\mathcal{A}^c}})\mathbf{1}_{T_{\mathcal{K}}>u}du\big)}{\mathbb{E}_{\vec{\partial}_{\mathcal{A}}} T_{\mathcal{A}}}.$$

At $u = 0$, the Markov process is at $X^\pi_{T_{\mathcal{A}^c}}$ which, as stated in Section 3.1, is distributed according to $p^\pi_{\vec{\partial}\mathcal{A}^c}$ when $X_0^\pi \sim p^\pi_{\vec{\partial}\mathcal{A}}$, as is the case above. This allows us to redefine the time origin of the Markov process at $u = 0$ and replace $\mathbb{E}_{\vec{\partial}\mathcal{A}}$ with $\mathbb{E}_{\vec{\partial}\mathcal{A}^c}$ to obtain

$$\mathbb{E}^\pi(\eta) = \frac{\mathbb{E}_{\vec{\partial}\mathcal{A}^c}\big(\int_0^{\infty} \eta(X_u^\pi)\mathbf{1}_{T_{\mathcal{K}}>u}du\big)}{\mathbb{E}_{\vec{\partial}_{\mathcal{A}}} T_{\mathcal{A}}}.$$

Expression (2) follows from replacing $u \to t$ and interchanging the order of the integral and the expectation, and expression (3) is directly obtained from (2) by conditioning on the event $\mathbf{1}_{T_{\mathcal{K}}>t}$. □

An estimator $\hat{\mathbb{E}}^\pi(\eta)$ of (3) is constructed by estimating each function inside the integral, as follows: the denominator of $g(t)$, $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_{\mathcal{A}})$, is estimated using regular Monte-Carlo from observations of the stopping time $T_{\mathcal{A}}$ coming from the simulation of X_t^π, $f_\eta(t)$ and the numerator of $g(t)$, $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)$, are estimated from the simulation of the FV N-particle system driven by X_t^π with absorption set $\mathcal{A}$. The estimation details are given in Appendix A.

3.3 Bound on the estimation error

It has been proved that for finite state spaces [8, 9], uniform in time propagation of chaos holds for Fleming-Viot particle systems. We let $m(\cdot, \xi) : \mathcal{A}^c \to [0, 1]$ denote the empirical distribution of the N particles with positions described by vector ξ, defined as the empirical mean $m(x, \xi) \doteq \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\xi(i)=x}, \forall x \in \mathcal{A}^c$. It then follows that if ν is a probability measure on $\mathcal{S}$, then, under assumption of proper initialization, [8, Theorem 1.4] shows the following bound on the speed of convergence w.r.t. the number of particles N of the empirical mean $m(., \xi_t^\nu)$ towards ϕ_t^ν:

$$\sup_{x\in\mathcal{S}}\sup_{t\geq 0}\mathbb{E}\Big|[m(x, \xi_t^\nu)] - \phi_t^\nu(x)\Big| \leq \frac{C_{\text{FV}}}{\sqrt{N}}, \qquad (4)$$

where C_{FV} is a positive constant depending on the characteristics of the driving process.

Using this result, we can show the error bound stated in the following theorem for the estimator $\hat{\mathbb{E}}^\pi(\eta)$ of (3) described at the end of Section 3.2. The theorem is valid in an idealized case where the simulation used to estimate the denominator of $g(t)$, $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_{\mathcal{A}})$, is started according to $p^\pi_{\vec{\partial}\mathcal{A}}$, and the Fleming-Viot simulation used to estimate $f_\eta(t)$ and the numerator of $g(t)$, $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)$, is started from i.i.d. samples of $p^\pi_{\vec{\partial}\mathcal{A}^c}$.

Theorem 2. Assume that we start the simulation of X_t^π for the estimator of $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_{\mathcal{A}})$ according to the distribution $p^\pi_{\vec{\partial}\mathcal{A}}$, that M return cycles to $\mathcal{A}$ under stationarity are observed during that simulation, and that we compute the estimator of f_η in (3) and of $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)$ using the FV particle system started at the positions of N i.i.d samples from $p^\pi_{\vec{\partial}\mathcal{A}^c}$. Let η be a bounded state function such that $\eta(x) = 0$ for $x \in \mathcal{A}$. Then, there exists a constant $C > 0$ such that

$$\mathbb{E}\left|\hat{\mathbb{E}}^\pi(\eta) - \mathbb{E}^\pi(\eta)\right| \leq C\Big(\frac{1}{\sqrt{M}} + \frac{1}{\sqrt{N}}\Big).$$

The proof is given in Appendix B.

Remark 1. Note that the estimator $\hat{\mathbb{E}}^\pi(\eta)$ is in general a biased estimator of $\mathbb{E}^\pi(\eta)$.

Remark 2. Theorem 2 ensures that the estimation error is of the same order as Monte-Carlo, which gives minimal guarantees for FV. However, the estimation error should not be our only focus to evaluate the difference between FV and Monte-Carlo, especially when the ultimate goal is the convergence of a reinforcement learning algorithm. Indeed, in the control problem, a noisy but still informative signal might be very useful compared to no signal at all. Our main idea when replacing MC by FV is to trade the observation of a very rare event in the original problem by the observation of a more common event for FV particles. Although a fully rigorous analysis of the probability to observe a non-zero reward is out of the scope of this paper, we can give rough estimations using the results of [8, 12]. The dynamics of a tagged particle of the FV process converges when $N \to \infty$ to a one-dimensional Markov process having as stationary measure the quasi-stationary distribution ν_{QS} of the original process (see [12]), where

$$\nu_{QS}(B) = \lim_{t\to\infty} \phi_t^{\vec{\partial}\mathcal{A}^c}(B),$$

for any set of states B. If the state space is finite, this one dimensional process in turn converges in distribution exponentially fast to its stationary distribution, ν_{QS}. Hence, the probability of finding a non-zero signal (by visiting states in set $\mathcal{C}$) for the FV process in a finite time interval is of the order of $\nu_{QS}(\mathcal{C})$, which can be significantly larger than the original $p^\pi(\mathcal{C})$ since the dynamics of FV particles within the restricted state space $\mathcal{A}^c$ are the same as those of the original process X_t^π. Thus, FV allows a shift of the estimation target, from $\mathbb{E}^\pi(\eta)$ to $\mathbb{E}^\pi_{\nu_{QS}}(\eta)$, where the latter can be significantly larger than the former.

3.4 FVRL: Policy gradient learning with Fleming-Viot particle systems

In this subsection we show how the Fleming-Viot estimation introduced in Section 3.2 can be combined with the policy gradient theorem to solve optimal control problems in environments with sparse and rare rewards under the expected reward criterion. For example, there are many MDPs (see for instance [22, 21, 14, 3, 15]) whose optimal policy is known to be of

threshold-type (because of underlying monotonicity properties); when the rewards structure in these MDPs is sparse and rare, learning the optimal thresholds with policy gradient algorithms becomes very slow because the gradient is zero except in very few and rarely observed states. We provide here a general description of the FVRL method we propose to speed up learning in those types of scenarios, which we will later apply to two network optimisation problems described in Section 4: an $M/M/1/K$ queue and a loss network system serving two classes of jobs.

Let us consider the case in which an agent interacts with an environment (available in practice either through a simulator, an emulator, or through experience replay of historical data), with the aim of maximizing the expected reward. As customary in the literature, we let S_n, R_n and A_n denote the state, reward and action at the n-th discrete time step, respectively. We denote by π_θ the policy function parameterised by θ. We recall that we assume that a non-zero reward is accrued only in states in $\mathcal{C}$.

It follows from classical MDP theory [20] that the optimal policy for the expected reward criterion is also optimal for the state value (bias) function defined as $v^{\pi_\theta}(x) \doteq \mathbb{E}^{\pi_\theta}\left[\sum_{n=0}^{\infty}\left(R_n - \overline{R}^{\pi_\theta}\right) \mid S_0 = x\right]$, where $\overline{R}^{\pi_\theta} \doteq \lim_{T\to\infty}\frac{1}{T}\mathbb{E}^{\pi_\theta}\sum_{n=0}^{T} R_n$ is the expected reward. We also define the corresponding state-action value function by

$$Q^{\pi_\theta}(x,a) = \mathbb{E}^{\pi_\theta}\left[\sum_{n=0}^{\infty}\left(R_n - \overline{R}^{\pi_\theta}\right) \mid S_0 = x, A_0 = a\right]. \tag{5}$$

It thus follows that $v^{\pi_\theta}(x) = \sum_a \pi_\theta(a|x) Q^{\pi_\theta}(x,a)$. We seek to find the policy π_θ that maximizes the average state value, $v^{\pi_\theta} = \sum_x p^{\pi_\theta}(x) v^{\pi_\theta}(x) = \sum_x p^{\pi_\theta}(x)\sum_a \pi_\theta(a|x) Q^{\pi_\theta}(x,a)$[1].

We propose to use a gradient-based algorithm to learn the parameter θ that maximizes v^{π_θ}. Denoting by X the random variable associated to the stationary distribution under π, it follows from the Policy Gradient Theorem [24] that the gradient of v^{π_θ} can be written as

$$\begin{aligned}\nabla_\theta v^{\pi_\theta} &= \mathbb{E}^{\pi_\theta}[Q^{\pi_\theta}(X,a)\nabla_\theta \pi_\theta(a|X)] \\ &= \sum_{x\in\mathcal{S}} p^{\pi_\theta}(x)\sum_a Q^{\pi_\theta}(x,a)\nabla_\theta\pi_\theta(a|x) \end{aligned} \tag{6}$$

We note that the gradient to learn θ is a linear combination of the policy gradients weighted by the stationary probability p^{π_θ}. Considering the assumed sparse and rare structure of the rewards in the environment to control, it is critical to observe the rare states sufficiently often in order to obtain a non-zero estimation of the stationary probability at those states. Otherwise, any non-zero policy gradient $\nabla_\theta\pi_\theta(a|x)$ coming from these rarely observed states (which are the most informative for learning θ in the context of the threshold-type policies being discussed) will not contribute to the gradient of the average state value, thus making learning difficult.

In this context, the FV particle system can be leveraged to obtain informative gradients of the average state value through an accurate estimation of the stationary probability p^{π_θ}, in particular at the rarely observed states. The combination of FV and RL leads to the FVRL algorithm, which is used to estimate the gradient of the average state value and learn the optimum θ.

The FVRL algorithm is derived from the following two-step estimation of the gradient of the average state value:

1. For each x where the policy gradient is non-zero, estimate the function of interest η introduced in Section 2, here defined as $\eta(x) = \sum_a Q^{\pi_\theta}(x,a)\nabla_\theta\pi_\theta(a|x)$.
2. Given an estimate $\hat{\eta}$ of η, use the FV procedure to estimate its expectation: $\sum_{x\in\mathcal{S}} p^{\pi_\theta}(x)\hat{\eta}(x)$.

This means that, besides estimating p^{π_θ} with Fleming-Viot, we need to estimate η for every x where the policy gradient is not zero for at least one possible action a. Thus, an estimate of η is constructed from estimates of $Q(x,a)$ for all actions a whose policy gradient at x is not zero. Although these $Q(x,a)$ values can be estimated by simulating separate Markov processes X_t^π independently –each starting at one of the different (x,a) contributing to η– here we propose a method based on coupled trajectories that leverages the sum-to-one property of the policy over all actions at state x. More precisely, the estimation method is as follows: if we let A_x be the set of possible actions when the system is at state x, we run $|A_x|$ copies of the Markov process on the extended state-action space, each starting at $(x,a_i)_{i=1,\dots,|A_x|}$. When two such chains meet at the same state-action, they continue evolving together forever. Thus, after all chains meet, the contribution to $\sum_a \hat{Q}^{\pi_\theta}(x,a)\nabla_\theta\pi_\theta(a|x)$ is zero, because at that point all values contributing to $\hat{Q}^{\pi_\theta}(x,a)$ are the same, and the partial derivatives of the policy sum up to zero for any given θ. Therefore, chains need to be run only until they all meet. Also, because of the sum-to-one property of the policy, the contribution to $\hat{\eta}$ from the term $\overline{R}^{\pi_\theta}$ in (5) is zero and therefore its value does not need to be computed, only the observed rewards R_n need to be recorded.

The following proposition states that the FVRL algorithm converges with probability 1 to the optimum parameter θ^*.

Proposition 3. Given a continuously differentiable parameterisation of the average state value v^{π_θ}, the policy gradient FVRL algorithm converges with probability 1 to the optimum parameter θ^*.

The proof is a consequence of [13, Section 5.2, Theorem 2.1] and [4], that consider the convergence of stochastic approximation algorithms with bias, as is the case with the FV estimator used in the estimation of the gradient of the average state value.

4 Application to stochastic networks with blocking

In this section we apply the methodologies outlined in Sections 3.2 and 3.4 to two different systems:

[1] With the same arguments presented in [20], it can easily be proven that the optimal policy for the expected reward criterion is optimal, not only for the state value (bias) function, but also for the average state value (i.e. its expectation over all states).

1. an $M/M/1/K$ queue, which is a unidimensional state system that will help us illustrate the basic concepts;
2. a loss network system with R servers receiving two classes of jobs at exponentially distributed inter-arrival times served with exponentially distributed service times, which we will use to illustrate the application to a multi-dimensional state system.

In both cases, we assume there exists a small subset $\mathcal{C}$ of the state space where costs may be generated due to the rejection of an incoming job (blocking), and that there is no reward for job acceptance. For each system we will consider two goals:

a) the accurate estimation of the expected cost, $\mathbb{E}^\pi(\eta)$, defined in Section 2;

b) the efficient learning of the blocking sizes that minimize the expected cost.

To simplify the exposition and to be able to do theoretical computations against which results are compared, we fix the parameters of the underlying Markov process. Nevertheless, the method is designed to be applied in practice to systems with possibly unknown parameters, the only condition being that the system can be simulated or emulated.

In all the simulations run to present the results of the method, either to estimate the expected cost or to learn the optimum blocking sizes, a burn-in period of 10 system transitions is considered before assuming that the stationary regime has been reached for the computation of estimators. In addition, the stationary probability estimate $p^\pi(x)$ –appearing in the calculation of the expected cost and of the policy gradient– is computed only when the number of observed return cycles to the absorption set $\mathcal{A}$, used in the estimation of $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_\mathcal{A})$, is at least 5 once the burn-in period has been completed[2]. If this condition is not satisfied the estimate of the stationary probability $p^\pi(x)$ is set to 0, except for the $M/M/1/K$ case where it is left undefined, reducing the number of successful estimations.

The algorithms for the estimation and for the optimisation problems in both the $M/M/1/K$ queue system and the loss network are presented in Algorithms 1 and 2 in Appendix D, respectively.

In each application problem, the method's performance is compared with vanilla Monte-Carlo (MC) which is used as benchmark. In order to assure a fair comparison, the MC estimation is based on the same number of events as the FV estimation, and the MC learning is based on the same number of events per learning step as the average number of events per learning step observed in FVRL. The details of the estimation and learning processes by MC are described in each subsection below.

4.1 M/M/1/K system

We start by defining the Markov process X_t^π that, under policy π, describes the dynamics of the $M/M/1/K$ system, namely the queue size that measures the number x of jobs waiting to be served in the buffer at a given time including the job being served. Jobs arrive following a Poisson process and their service times are assumed exponentially and independently distributed. We consider that the policy π belongs to the family of deterministic accept/reject policies that reject an incoming job at just one state, when $X_t^\pi = K$, in which case a unitary rejection cost is accrued.

The process is thus a continuous-time discrete-state stochastic process living in $\{0, 1, \dots, K\}$ with upward jump rate λ from any state x to $x+1$ (except at $x = K$), and downward jump rate μ from x to $x-1$ (except at $x = 0$). We denote by $\rho \doteq \lambda/\mu$ the load of the system which is typically smaller than 1 in real applications. Thus, the system can be represented by an MDP $(\mathcal{S} = \{x : 0 \le x \le K\}, A = \{0, 1\}, q = \{\lambda, \mu\}, \mathcal{R} = r(x, a) = \mathbf{1}_{\{x=K, a=0\}})$ where actions 0 and 1 represent "reject" and "accept" an incoming job, respectively.

All presented results correspond to an $M/M/1/K$ system that serves, at rate $\mu = 1$, jobs arriving at rate $\lambda = 0.7$.

4.1.1 Estimation of the expected cost

We apply the methodology outlined in Section 3.2 to efficiently estimate the expected cost when blocking is a rare event.

Since the rejection cost is 1, estimating the expected cost is tantamount to estimating the rejection probability, which can be quite small depending on the system parameters. For example with $\rho = 0.5$ and $K = 20$, the rejection probability becomes of the order of 10^{-6}, a result obtained from the PASTA property of Poisson arrivals and the calculation of the stationary probability where rejection occurs, which is $\mathbb{P}(X_t^\pi = K) = \frac{(1-\rho)\rho^K}{1-\rho^{K+1}}$ whenever $\rho < 1$.

In order to estimate the blocking probability using Fleming-Viot, we define the function η introduced in Section 2 as $\eta(x) = \mathbf{1}_{\{x=K\}}$, which is sparse and non-zero in the single-state set $\mathcal{C} = \{K\}$. As a consequence, the expectation $\mathbb{E}^\pi(\eta)$ in (3) becomes the blocking probability, which can be written as

$$\mathbb{E}^\pi(\eta) = p^\pi(K) = \frac{\int_0^\infty \phi_t^J(K)\mathbb{P}_J(T_\mathcal{K} > t)dt}{\mathbb{E}_{J-1}T_\mathcal{A}}, \qquad (7)$$

where we have used that any set $\mathcal{A} = \{0, 1, \dots, J-1\}$ with $J \le K$ is a valid absorption set, making $\vec{\partial}A$ and $\vec{\partial}A^c$ two single-state entrance boundary sets equal to $\{J-1\}$ and $\{J\}$, respectively, which makes it possible to simplify (3) into (7).

According to Theorem 2, the estimator of $\mathbb{E}^\pi(\eta)$ converges to its true value as both the number N of particles of the FV system, and the number M of return cycles to $\mathcal{A}$ increase. For simulation purposes, and since the number M of return cycles to $\mathcal{A}$ is random, we will refer to the number of arrival events T (which directly impacts M) when defining the hyperparameters of experiments. The three quantities in (7) contributing to the blocking probability are estimated following the steps described in Algorithm 1 in Appendix D, which implements the estimation methodology described in Appendix A.

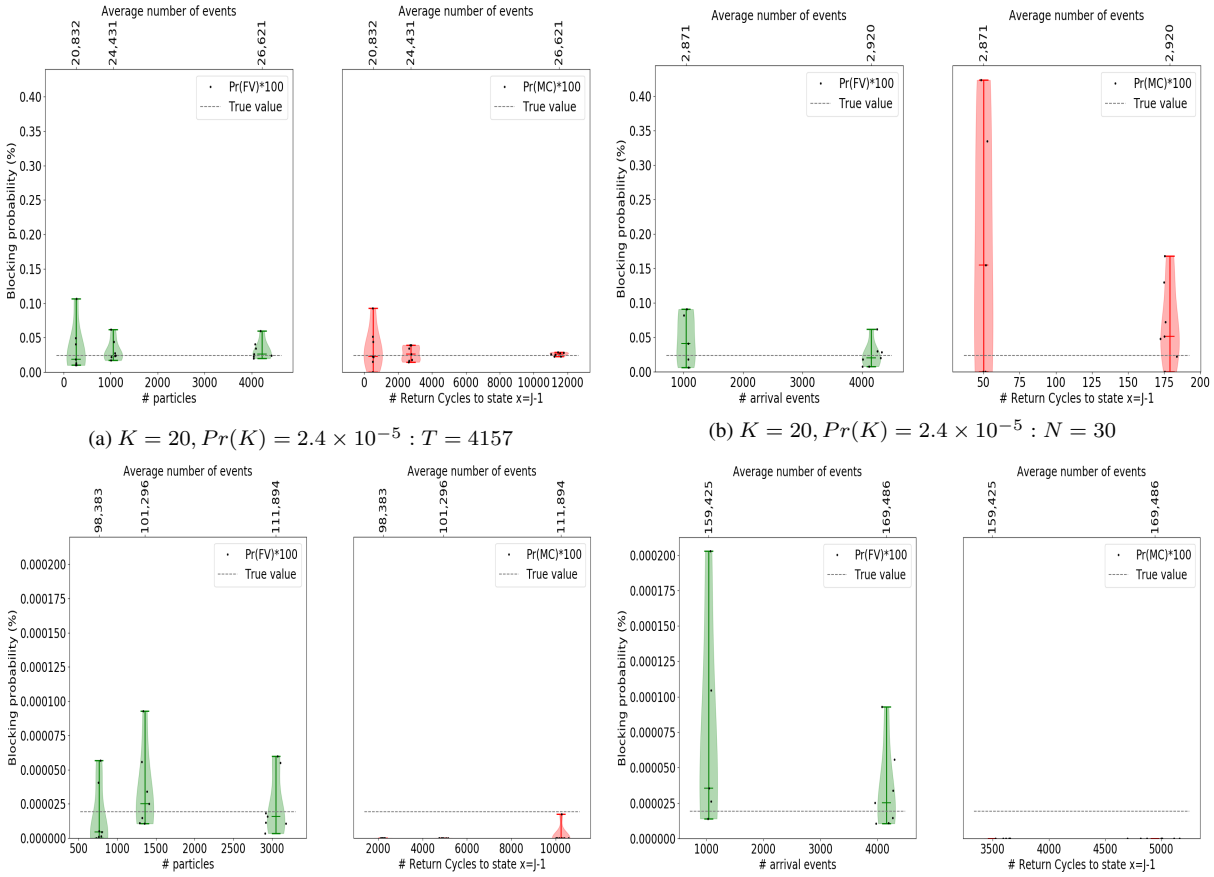
Remark 3. There is a trade-off between choosing a small or a large value of J, the state that defines the size of the absorption set $\mathcal{A}$: for smaller J, the return times to $\mathcal{A}$ will be smaller,

[2] The return cycles to $\mathcal{A}$ are measured from the first time the system enters $\mathcal{A}$ after the burn-in period has been overcome.

requiring fewer arrival events T for an accurate estimation of the denominator in (7), but at the same time visiting the rare blocking state K will be rarer, requiring a larger number of particles N for an accurate estimation of the numerator in (7). The opposite is true for larger values of J. A detailed analysis on this trade-off in terms of appropriately choosing N and T for accurate estimations of the numerator and the denominator in (7) is presented in Appendix C.

We now study the convergence of the FV estimator as either N or T increases, and compare it with the benchmark MC estimator, obtained from a direct application of expression (1), i.e. as the fraction of the time spent at state K and the total time of return cycles to the initial state $x = J - 1$ observed during the simulation. To guarantee a fair comparison between the two methods, we start the simulation at $x = J - 1$, so that both methods start at the same "distance" from the blocking state K, and let the simulation run until the same number of events observed in the FV estimator is reached.

Figure 2 compares the convergence of the FV estimator with the MC estimator both as N increases and as T increases. We considered the cases $K = 20$ and $K = 40$, which are regarded to represent moderate and large capacities based on their blocking probabilities at the considered value for $\rho = 0.7$ of order 10^{-4} and 10^{-7}, respectively. The size J of the absorption set $\mathcal{A}$ is held fixed at $J = 12$.



(a) $K = 20, Pr(K) = 2.4 \times 10^{-5} : T = 4157$

(b) $K = 20, Pr(K) = 2.4 \times 10^{-5} : N = 30$

(c) $K = 40, Pr(K) = 1.9 \times 10^{-7} : T = 4157$

(d) $K = 40, Pr(K) = 1.9 \times 10^{-7} : N = 1356$

Figure 2: Violin plots showing the convergence properties of the Fleming-Viot (green) and Monte-Carlo (red) estimators of the blocking probability for an $M/M/1/K$ queue system with $\lambda = \rho = 0.7$. Left plots (a) and (c): convergence as the number of particles N increases with T fixed. Right plots (b) and (d): convergence as the number of arrival events T increases with N fixed. The absorption set size is $J = 12$ in all cases. For the left set of plots (a) and (c), the number of particles considered for the convergence analysis are $N = 264, 1055, 4220$ for $K = 20$, and $N = 763, 1356, 3051$ for $K = 40$. For the right set of plots (b) and (d), the number of arrival events considered for the convergence analysis are $T = 462, 1040$, for both $K = 20$ and $K = 40$. For more details on these choices, see Appendix C. The middle horizontal line in each violin plot represents the median. In the corresponding MC experiments, the average number of observed return cycles to the initial state $J - 1$ is used on the horizontal axis while the top horizontal axes show the average number of events observed in the experiments run in each set, which by design coincide between each paired FV-MC execution, as described in the text.

We observe the following in terms of convergence of the FV and MC estimators to the true blocking probability, which in the plots is represented by a horizontal gray line:

1. Both the FV and the MC estimators converge to the true blocking probability when $K = 20$ (Figures 2(a,b)). MC presents a smaller variability than FV in the convergence analysis with N (Figure 2(a)) but a larger variability than FV in the convergence analysis with T (Figure 2(b)). This is due to the fact that Figure 2(a) is obtained from simulations whose events largely outnumber the events in Figure 2(b) by as much as 10 times ($\sim 20,000$ vs. $\sim 2,000$), which allows MC to observe the blocking event at $x = K$ frequently enough for an accurate estimation of the blocking probability. On the other hand, Figure 2(b) tells us that a much smaller number of events ($\sim 2,000$) is sufficient for FV to estimate the blocking probability accurately enough, but is not sufficient for MC. This demonstrates the higher efficiency of FV than MC in discovering the rare event at K.
2. When $K = 40$ (Figures 2(c,d)) the MC estimator basically fails as it almost never observes the blocking state K.
3. Increasing N has a larger impact in increasing the FV estimator's computational complexity than increasing T. This conclusion is obtained by observing that the number of events at the leftmost violin plot in Figure 2(a) ($\sim 20,000$ events), where $N \sim 4,000$ and $T \sim 100$, is about 10 times larger than the number of events at the rightmost violin plot in Figure 2(b) ($\sim 3,000$ events), where $N \sim 100$ and $T \sim 4,000$ (i.e. the N and T values are switched w.r.t. the previous case).

4.1.2 Learning the optimum blocking size using FVRL

We illustrate the FVRL learning algorithm by defining a cost function of rejecting an incoming job that is exponentially increasing with the queue size at the time of the job arrival. Note that the blocking cost needs to be exponentially increasing with the queue size in order to obtain an optimisation problem with a non-trivial solution (i.e. where the optimum K is finite). This is due to the fact that the stationary probability is exponentially decreasing with the queue size, as mentioned at the end of Section 4.1, and such exponentially increasing cost function guarantees that the expected cost to minimize is a convex function of K. More precisely, given the queue size x, the cost function is defined as $r(x, a) = B(1 + b^{x-x_{ref}})\mathbf{1}_{\{a=0\}}$, where B, b and x_{ref} are positive constants which, for a given load ρ, are carefully chosen to make the expected cost to minimize a convex function. From [14] we know that the policy optimising such convex cost functions is of threshold type, that is, there exists a state $0 < K < \infty$ such that $a = 1$ is optimal for all $x < K$, and $a = 0$ is optimal for $x = K$. The value of x_{ref} is the reference queue size that is instrumental in defining the optimum threshold, K^*, which is the closest integer to x_{ref}.

We now use the policy gradient methodology presented in Section 3.4 to learn the optimum threshold K^*. Following [16] and using the framework of Semi-Markov Decision Processes (SMDP) to describe the continuous-time Markov process X_t^π by a discrete-time process X_n^π [19, 16], we propose a parameterised acceptance policy $\pi(a = 1|x)$ that is a linear step function of the state x, which is deterministic for x outside the interval $(\theta, \theta + 1)$ and decreases linearly from 1 to 0 within the interval. That is, the acceptance policy parameterised by the positive real-valued θ, is defined as:

$$\pi_\theta(a = 1|x) = \begin{cases} 1 & \text{if } x \leq \theta, \\ x - \theta + 1 & \text{if } \theta < x < \theta + 1, \\ 0 & \text{if } x \geq \theta + 1. \end{cases} \quad (8)$$

Note that the policy is deterministic for integer-valued θ, in which case the blocking size is $K = \theta + 1$. Otherwise, K is defined as $\lceil \theta \rceil + 1$.

We use a gradient descent algorithm to learn the optimum parameter θ that minimizes the average state value v^{π_θ}, where "value" here is thought of as cost. Using expression (6), the gradient of v^{π_θ} becomes

$$\frac{\partial v^{\pi_\theta}}{\partial \theta} = p^{\pi_\theta}(K - 1)\left[Q^{\pi_\theta}(K - 1, 1) - Q^{\pi_\theta}(K - 1, 0)\right], \quad (9)$$

where $K - 1$ is the smallest integer that is larger than or equal to θ. We observe that this parameterisation leads, as expected, to gradients being 0 for $x \neq K - 1$, that is, the policy gradient is sparse. We note also that the gradient is discontinuous at θ and $\theta + 1$, making the assumptions of Proposition 3 not fully satisfied. However, these two points have measure zero and therefore, with probability 1, no discontinuity is observed[3].

Details of the learning algorithm are given in Algorithm 2 in Appendix D.

To illustrate the FVRL algorithm, we consider an MDP with the following characteristics: the system load is $\rho = 0.7$, the blocking cost function $r(x, a)$ is defined with the parameters $b = 3$ ($> 1/\rho$ so that the expected cost function is convex), $B = 5$, and $x_{ref} = 18$, giving $K^* = 18$.

The setup of the learning experiments is as follows: we choose the value $(x_{ref} + 10)$ for the initial blocking size guess, so that, already at the onset, blocking occurs rarely. Since the value of K is no longer fixed (as was the case when estimating the blocking probability) but is now learned by the algorithm, it is not possible to choose a fixed value J for the size of the absorption set $\mathcal{A}$. Instead, we consider a fixed J/K fraction that adapts J to each value of K at the start of each learning step. In order to experiment with different sizes of the absorption set $\mathcal{A}$, we consider two different scenarios: $J = \lceil 0.3K \rceil$ and $J = \lceil 0.5K \rceil$. Following the conclusions of the choice of J, N and T and their impact in the FV estimator accuracy outlined in Appendix C, for each value of J, we adjust N and T at each learning step to obtain approximate expected relative errors of $\epsilon_\phi \sim 100\%$ and $\epsilon_{ET} \sim 20\%$, respectively for

[3] A special case occurs when θ is integer, in which case the discontinuities would be observed with non-zero probability in the gradient descent algorithm under the following scenario: an integer value is chosen for the initial guess of θ, and integer-valued clipping (e.g. to ± 1) is used for the next θ estimated by the algorithm. This problem is solved by simply not choosing an integer-valued initial guess of θ.

the estimations of $\phi_t^J(K-1)$ and $\mathbb{E}_{J-1}(T_{\mathcal{A}})$, i.e. we set the approximate expected error for $\hat{\phi}_t^J(K-1)$ much larger than the approximate expected error for $\hat{\mathbb{E}}_{J-1}(T_{\mathcal{A}})$. For comparison purposes and to further check the conclusions of the error analysis in Appendix C, we also considered a scenario where the pre-defined relative errors are inverted, namely $\epsilon_\phi \sim 20\%$ and $\epsilon_{ET} \sim 100\%$.

For each setup we ran the FVRL policy learner on 100 learning steps. At each learning step, as long as a positive estimate of the stationary probability $p^{\pi_\theta}(K-1)$ has been obtained, we also estimate the sparse $\eta(x)$ function defined in Section 3.4 –which in this case is simply the difference between two Q values as seen in (9)– using the coupling procedure described in Section 3.4 and allowing up to 250 arrival events until mixing is observed in each of 100 replications. The value of $\hat{\eta}(x)$ is finally computed as the average Q-difference over these replications[4] and multiplied with the stationary probability estimate $\hat{p}^{\pi_\theta}(K-1)$. Parameter θ is then updated by gradient descent using a constant learning rate of $\alpha = 10$, and the number of observed events at each learning step is recorded. The estimated optimum threshold is set to $\hat{K}^* = \text{round}(\hat{\theta}^*) + 1$, where $\hat{\theta}^*$ is the value of the θ parameter obtained after the last learning step.[5].

The benchmark MC learning algorithm uses the same policy gradient approach as FVRL with the only difference that it estimates the probability $p^{\pi_\theta}(K-1)$ in expression (9) using Monte-Carlo instead of Fleming-Viot, i.e. based on a single trajectory of X_t^π as the ratio between the continuous time that the system spends at $K-1$ and the total return cycle time to $J-1$ under stationarity[6], where J is defined as a function of K as in FVRL. Each replication of the MC learner is started at $J-1$ and is stopped when the average number of observed events over all learning steps in the respective FVRL replication is observed. These two conditions allow a fair comparison between the two methods.

The results of the above procedure run on 20 replications are shown in Figure 3. We see that FVRL clearly outperforms the benchmark MC learning in three out of the four scenarios considered, depicted in Figures 3(a), 3(b) and 3(d). In the scenario depicted in Figure 3(c), learning by the two methods is similar because the average number of events is about 10 times larger than the one observed in the other three scenarios ($\sim 40,000$ vs. $\sim 5,000$), which allows MC to also observe the rare blocking state and thus learn almost as fast as FVRL. This is the same situation previously observed about the probability estimation results presented in Figure 2.

Of the two setups, $\epsilon_\phi \sim 100\%$ and $\epsilon_{ET} \sim 20\%$ used for Figures 3(a) and 3(b), and $\epsilon_\phi \sim 20\%$ and $\epsilon_{ET} \sim 100\%$ used for Figures 3(c) and 3(d), the setup that is both the safest and the least computationally intensive is the former. This is consistent with the conclusions of the analysis of the choice of J, N and T in FV presented in Appendix C, namely that, for an accurate estimation of the stationary probability, more importance should be given to achieve an accurate estimation of $\mathbb{E}_{J-1}(T_{\mathcal{A}})$ than an accurate estimation of $\phi_t^J(K)$. In terms of safety we note that, for the least convenient setup of $\epsilon_\phi = 20\%$ and $\epsilon_{ET} = 100\%$ of Figure 3(d), in one replication the θ parameter suddenly increases from 28 to 140 (not shown but made apparent by the mean learning curve being significantly above the median learning curve). This overshoot impedes further learning because the blocking probability becomes extremely small at $K = 140$. This clearly illustrates the risks of choosing a too small T value (associated to the large error $\epsilon_{ET} = 100\%$) which may considerably overestimate the blocking probability (due to an underestimation of $\mathbb{E}_{J-1}(T_{\mathcal{A}})$) and generate such out-of-control excursion coming from a noisy estimation of the Q-difference contributing to the gradient expression.

4.2 Loss network system

Consider now a loss network that processes jobs of I different classes which, as in the $M/M/1/K$ system, arrive following independent Poisson processes and are served with independently exponentially distributed times by one of R available servers. We are interested in analyzing the number of jobs of each class being served by the system at a given time t, $\mathbf{X}_t^\pi = (X_{t,1}^\pi, ..., X_{t,I}^\pi)$, whose dynamics is governed by the system characteristics and the policy π being applied. As before, we also consider a policy π that belongs to the family of deterministic accept/reject policies that reject an incoming job of a given class when the system is already serving a predefined number of jobs of that class, in which case a class-dependent rejection cost is accrued.

Let us denote by λ_i, μ_i, C_i, K_i the job arrival rate, the service rate, the rejection cost, and the blocking size of each job class $i = 1, \dots, I$. Thus, the system can be represented by an MDP ($\mathcal{S} = \{\mathbf{X} \in \mathbb{R}^I : 0 \leq X_i \leq K_i, i = 1, \dots, I \text{ s.t. } \mathbf{X}^T\mathbf{1} \leq R, \}, A = \{0, 1\}, q = \{\lambda_1, \dots, \lambda_I; \mu_1, \dots, \mu_I\}, \mathcal{R} = r(\mathbf{x}, a) = \sum_{i=1}^I C_i \mathbf{1}_{\{\mathbf{x} \in \mathcal{S}_i, \mathcal{L}_i, a=0\}}$), where S_i is the set of states where blocking occurs following an i-class job arrival, an event that is denoted by $\mathcal{L}_i$.

For the purposes of simplifying illustration, in our experiments we consider the smallest multi-class loss network, one that serves just two classes of job.

4.2.1 Estimation of the expected cost

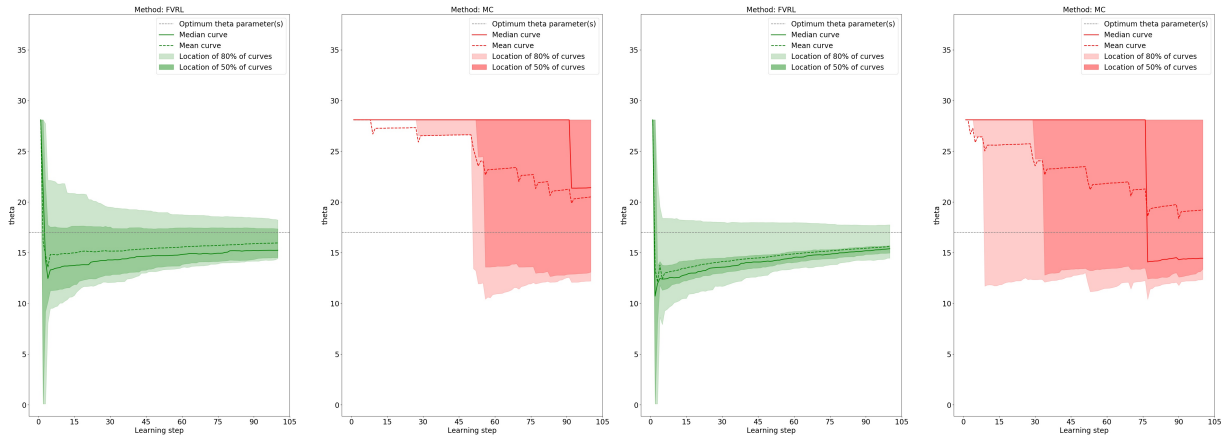
Contrary to the $M/M/1/K$ case, estimating the expected cost in the loss network system is not equivalent to estimating the blocking probability, because blocking occurs at more than one state whose cost in general depends on the class of the arriving job being rejected.

The expression of the expected cost $\mathbb{E}^\pi(\mathcal{R})$ under the

[4] The number of replications on which the Q-difference is averaged may be less than 100 because a replication is excluded from the average when mixing does not occur, but this occurs very rarely.

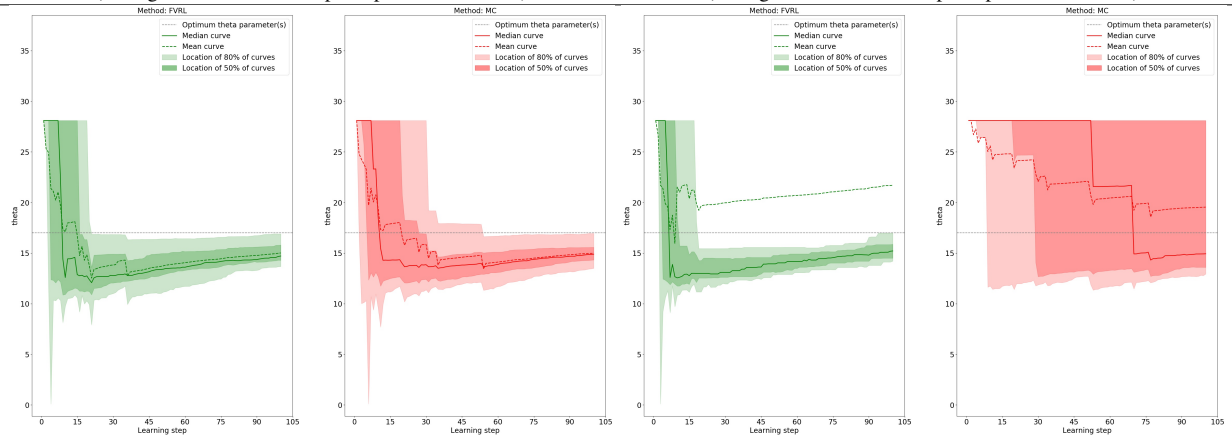
[5] Note that the estimated optimum threshold is not set to $\lceil \hat{\theta}^* \rceil + 1$, as in the definition of the parameterised policy, so that $\hat{K}^*$ is more naturally chosen to be e.g. 5 when $\hat{\theta}^* = 4.01$ rather than 6.

[6] Stationary is assumed after the burn-in period whose value is defined in Section 4.



(a) $\hat{\phi} \sim 100\%, \hat{\mathbb{E}}(T_{\mathcal{A}}) \sim 20\%, J/K = 0.3$
(average number of events per replication is 4,000)

(b) $\hat{\phi} \sim 100\%, \hat{\mathbb{E}}(T_{\mathcal{A}}) \sim 20\%, J/K = 0.5$
(average number of events per replication is 5,000)



(c) $\hat{\phi} \sim 20\%, \hat{\mathbb{E}}(T_{\mathcal{A}}) \sim 100\%, J/K = 0.3$
(average number of events per replication is 40,000)

(d) $\hat{\phi} \sim 20\%, \hat{\mathbb{E}}(T_{\mathcal{A}}) \sim 100\%, J/K = 0.5$
(average number of events per replication is 5,000)

Figure 3: Comparison of FVRL learning (left subplot in green) with Monte-Carlo learning (right subplot in red) of the optimum parameter $\theta^* = 17$ (horizontal dashed gray line) of the parameterised acceptance policy π_θ of incoming jobs to the $M/M/1/K$ queue system with $\lambda = \rho = 0.7$. Each plot caption indicates the following learning settings for the FVRL approach: the approximate expected relative error wished for the estimation of $\phi_t(K)$ (which defines the number of particles N at each learning step), the approximate expected relative error wished for the estimation of $\mathbb{E}(T_{\mathcal{A}})$ (which defines the number of arrival events T at each learning step), and the fraction J/K defining the size of the absorption set. In all cases the learning parameter is kept constant at $\alpha = 10$, and the initial θ guess is 28.1, which defines a "large" blocking size of $K = 30$ whose small stationary probability ($\sim 10^{-6}$) makes blocking a rare event already at the onset of learning.

threshold policy π and rewards $\mathcal{R}$ is obtained from the total probability theorem applied on all possible job arrival classes and all possible accept/reject actions on those arrivals, as:

$$\mathbb{E}^{\pi}(\mathcal{R}) = \sum_{x \in \mathcal{S}} p^{\pi}(\mathbf{x}) c(\mathbf{x}), \tag{10}$$

where $p^{\pi}(\mathbf{x}) \doteq Pr(\mathbf{X}_n^{\pi} = \mathbf{x})$ is the stationary probability that the SMDP $\mathbf{X}_n^{\pi}$ (defined in Section 4.1.2) is at state $\mathbf{x}$, and $c(\mathbf{x}) = \sum_{i=1}^{I} C_i \lambda_i / \Lambda \left[\mathbf{1}_{\{\sum_{j=1}^{I} x_j = R\}} + \mathbf{1}_{\{\sum_{j=1}^{I} x_j < R\}} \mathbf{1}_{\{x_i = K_i\}} \right]$ is the expected cost of rejection at $\mathbf{x}$ over all possible arriving job classes, where $\Lambda \doteq \sum_{i=1}^{I} \lambda_i$ is the total job arrival rate.

The true expected cost is computed from (10) using the stationary probability of the stochastic knapsack as $p^{\pi}(\mathbf{x})$ [21, Chapter 4] , against which the estimated expected cost is compared.

The convergence of the estimated expected cost to its true value is analyzed via the violin plots shown in Figure 4, where, following the conclusions about the choice of J, N and T for the $M/M/1/K$ queue system described in Appendix C, is analyzed in terms of increasing T (as opposed to increasing N), since a large value of T is more crucial than a large value of N for an accurate estimation. In order to analyze the impact on the estimation accuracy of the distribution of the start states, two scenarios are considered in this analysis: (a) one where the N initial states for the FV simulation used to estimate $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)$ and $\phi_t^{\vec{\partial}\mathcal{A}^c}$ are chosen uniformly at random out of the states in $\vec{\partial}\mathcal{A}^c$, and (b) one where the initial states are chosen according to the stationary distribution of the states in $\vec{\partial}\mathcal{A}^c$, which is known for the loss network considered.

The latter is still not the correct distribution to use for the initial states, but is a better approximation than the uniform to the actual distribution that should be used according to Theorem 2, namely the *entrance* stationary distributions of the states in $\vec{\partial}\mathcal{A}^c$. In practice, when the stationary distribution of the states in $\vec{\partial}\mathcal{A}^c$ is unknown, one should use the entrance stationary distribution of the states in $\vec{\partial}\mathcal{A}^c$, estimated using the procedure described in Appendix A and implemented by Algorithm 1, which is based on the single simulation of the Markov process that estimates $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_{\mathcal{A}})$. Note that, as stated in Appendix A, the initial state for the estimation of $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_{\mathcal{A}})$ is always chosen uniformly at random among the states in $\vec{\partial}\mathcal{A}$, and after that the burn-in period is used so that the system is closer to stationarity before collecting the data for the estimation process.

Finally, to guarantee a fair comparison with the benchmark, the simulation for the MC estimator is started at a uniformly randomly chosen state in $\vec{\partial}\mathcal{A}$ –so that both methods start at a similar "distance" from the set of blocking states– and is let run until the number of events of the FV simulation is observed.

A loss network with the following characteristics was considered for the estimation of the expected cost: capacity $R = 6$ servers, $\lambda = [1, 5]$, $\mu = [3.33, 50.0]$, hence $\rho = [0.3, 0.1]$, $C = [2.5 \times 10^3, 4.9 \times 10^6]$, $K = [4, 6]$. The choice of the rejection costs allows us to illustrate the benefits of the FV approach over MC as it makes a few of the smaller probability cost-generating states as important as those with larger probability in terms of their contribution to the system's expected cost. These contributions are listed in Table 1, together with their respective stationary probabilities which have been computed following the product form of the stationary probability of the stochastic knapsack described in [21, chapter 4]. We observe that the top three states in terms of probability contribute to about 65% of the expected cost, while the bottom four, with probability smaller than 10^{-6}, contribute to as much as 35%, therefore obtaining accurate estimates of the probability of a few of those smaller probability states is important for an accurate estimation of the expected cost.

State $\mathbf{x}$	**Prob.** $p^\pi(\mathbf{x})$	**Expected cost** $c(\mathbf{x})$	$p^\pi(\mathbf{x})c(\mathbf{x})$	**% Exp. cost** $\mathbb{E}^\pi(\mathcal{R})$
$[4, 0]$	2.2×10^{-4}	0.416×10^3	0.094	1.3%
$[4, 1]$	2.3×10^{-5}	0.416×10^3	0.009	0.2%
$[4, 2]$	1.1×10^{-6}	4084×10^3	4.525	62.7%
$[3, 3]$	5.0×10^{-7}	4084×10^3	2.011	27.8%
$[2, 4]$	1.3×10^{-7}	4084×10^3	0.503	7.0%
$[1, 5]$	1.7×10^{-8}	4084×10^3	0.067	0.9%
$[0, 6]$	9.3×10^{-10}	4084×10^3	0.004	0.1%
Total			7.214	100.0%

Table 1: Contribution to the expected cost $\mathbb{E}^\pi(\mathcal{R})$ by each blocking state in the loss network system with $R = 6$, $\lambda = [1, 5]$, $\mu = [3.33, 50.0]$ ($\rho = [0.3, 0.1]$), $C = [2.5 \times 10^3, 4.9 \times 10^6]$, blocking sizes $K = [4, 6]$ defining policy π, sorted by decreasing probability $p^\pi(\mathbf{x})$.

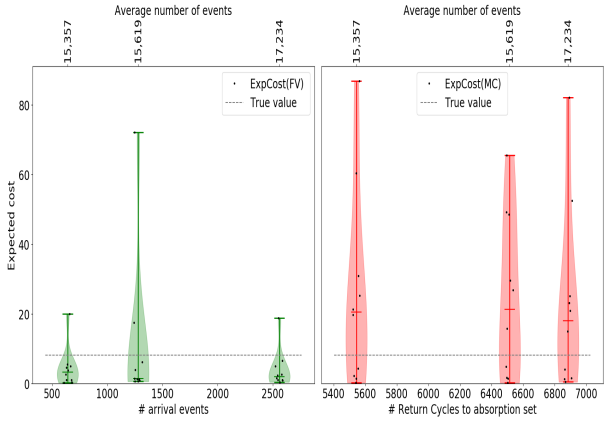
Figure 4 compares the convergence of the FV estimator with the MC estimator of the expected cost of the system as the number of arrival events T increases, while keeping a fixed N set to 500 and an absorption set $\mathcal{A}$ with size $J = [1, 2]$ in each job class dimension. We observe:

- a consistently smaller variance of the FV estimator compared to the MC estimator.
- a higher accuracy of the FV estimator (although with larger variance) when the start state of the FV simulation is chosen following the stationary distribution of states in $\vec{\partial}\mathcal{A}^c$ compared to the uniformly random choice case. In the latter case (Figure 4(a)) the FV estimates tend to underestimate the expected cost, as made apparent by the median value represented by the horizontal line in the middle of each violin plot.
- a tendency of the MC estimator to overestimate the expected cost.

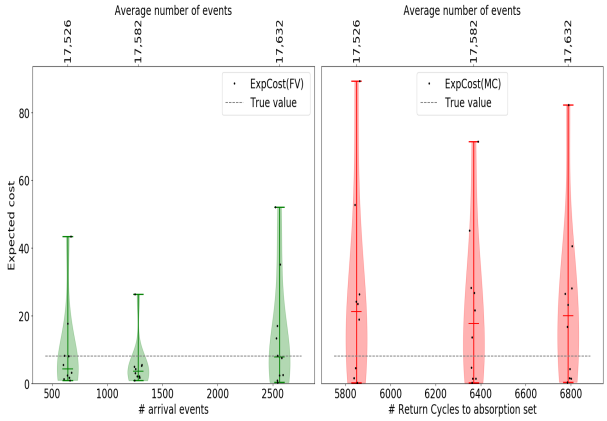
4.2.2 Learning the optimum blocking sizes using FVRL

The FVRL algorithm of the optimum blocking sizes $\mathbf{K}^* \in \mathbb{N}^I$ of a loss network serving I jobs classes learns the optimum θ_i of I parameterised acceptance policies, each of the form (8). For each θ_i observed during learning, the deterministic blocking size K_i of each policy is defined as $K_i = \lceil \theta_i \rceil + 1$, but, as in the $M/M/1/K$ case, the estimated optimum blocking size obtained at the end of the learning process is defined as $\hat{K}_i^* = \text{round}(\hat{\theta}_i^*) + 1$.

Upon arrival of a job of class i, the respective π_{θ_i} acceptance policy is applied, making the system's acceptance policy equal to $\pi_\theta(a = 1|\mathbf{x}) = \sum_{i=1}^I \pi_{\theta_i}(a = 1|x_i)\mathbf{1}_{\mathcal{L}_i}$, where $\mathcal{L}_i$ denotes the arrival event of a class-i job. Its derivative w.r.t. θ_i is non-zero only at states $\mathbf{x}$ for which $x_i = K_i - 1$ as long as they satisfy the R-server constraint for a possible job acceptance, $\mathbf{x}^T\mathbf{1} < R$. At those states the derivative is equal to $+1$ for action $a = 1$ (accept) and -1 for action $a = 0$ (reject).



(a) The start state of each FV particle is chosen uniformly at random in the boundary set $\mathcal{A}^c$



(b) The start state of each FV particle is chosen following the stationary distribution of the states in the boundary set $\mathcal{A}^c$

Figure 4: Violin plots showing the convergence properties of the Fleming-Viot (left) and Monte-Carlo (right) estimators of the expected cost in a two-class loss network with $R = 6$ servers, $\lambda = [1, 5]$, $\rho = [0.3, 0.1]$, blocking at $K = [4, 6]$ with costs $C = [2.5 \times 10^3, 4.9 \times 10^6]$, as the number of arrival events T increases. The absorption set sizes by job class are $J = [1, 2]$ and the number of particles is $N = 500$. The middle horizontal line in each violin plot is the median estimated value among 10 replications. In the corresponding MC experiments, the average number of observed return cycles to the absorption set is used on the horizontal axis. Finally, the top horizontal axes show the average number of events observed in the experiments run in each set, which by design coincide between each paired FV-MC execution, as mentioned in the text.

Hence, the partial derivative of v^{π_θ} w.r.t. θ_i becomes:

$$\frac{\partial v^{\pi_\theta}}{\partial \theta_i} = \sum_{\mathbf{x} \in \mathcal{S}_i} p^{\pi_\theta}(\mathbf{x}) \left[Q^{\pi_\theta}(\mathbf{x}, 1) - Q^{\pi_\theta}(\mathbf{x}, 0) \right], \text{ for } i = 1, \ldots, I, \quad (11)$$

where $\mathcal{S}_i$ is the set of states $\mathbf{x}$ where $x_i = K_i - 1$ and $\mathbf{x}^T \mathbf{1} < R$.

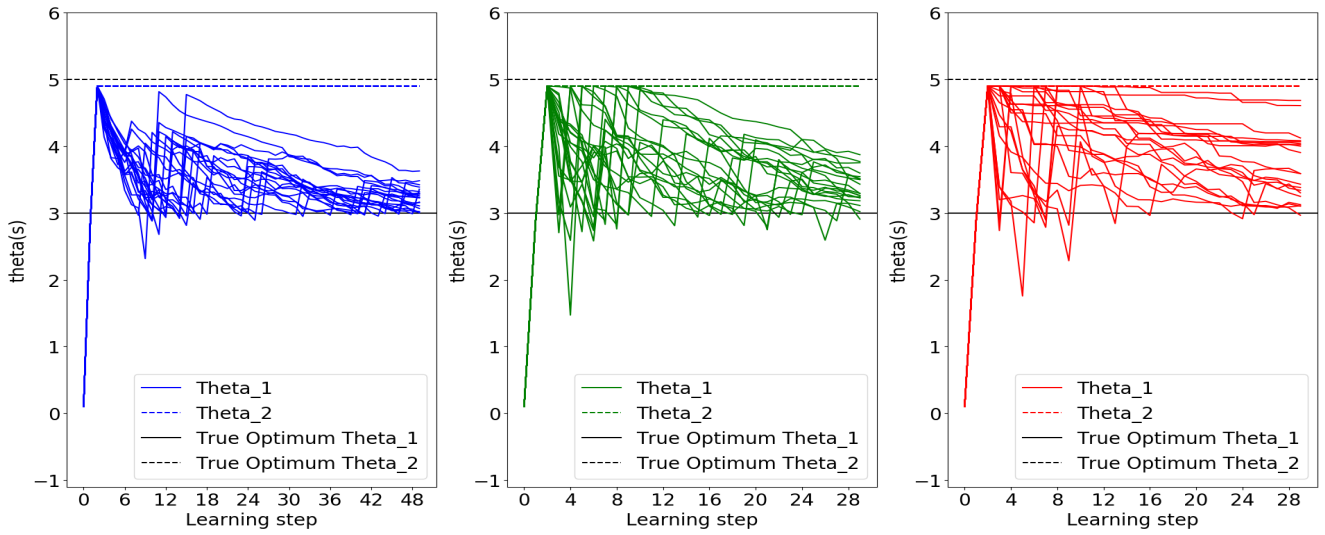
To illustrate the FVRL algorithm, we consider a two-job-class loss network with the following characteristics: $R = 6$ servers, job arrival rate by class $\lambda = [1, 5]$, service rate by class $\mu = [2.0, 16.67]$ (resulting in a load by class $\rho = [0.5, 0.3]$), and a blocking cost by class $C = [2 \times 10^3, 2 \times 10^4]$. For this choice of parameters, the optimum blocking sizes of the threshold policy are equal to $K^* = [4, 6]$ (i.e. $\theta^* = [3, 5]$), which will be estimated with FVRL.

The setup of the learning experiments is as follows: we set the initial blocking size guess to $K = [0.1, 0.1]$, and the following simulation parameters for the estimation of the stationary probabilities $p^{\pi_\theta}(\mathbf{x})$ for each $\mathbf{x} \in \mathcal{S}_i$: $J/K = 0.5$ for both classes, the number of FV particles $N = 100$, and the number of arrival events $T = 500$. The start state of each FV particle is chosen following the known stationary distribution of the states in $\vec{\partial}\mathcal{A}^{c(7)}$. The estimation of the gradient in (11) is completed by estimating the Q-difference at each $\mathbf{x}$ following the same coupling procedure used for the $M/M/1/K$ case. The learning rate α is set initially to 1.0 and is then decreased inversely proportional to the learning step.

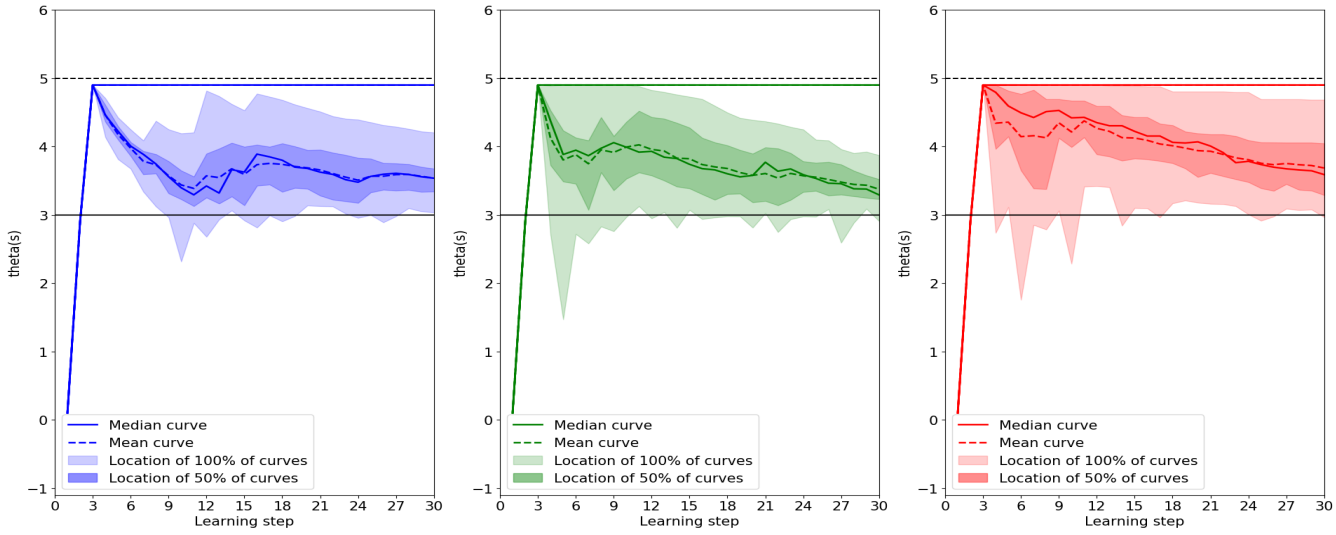
The results of the above procedure run on 20 replications are shown in Figure 5 and described in the caption. We see that FVRL outperforms MC learning in terms of median learning speed of parameter θ_1 with non-trivial optimum, and presents a smaller path variability as well. On the other hand, the FVRL and MC average learning speeds are similar and are a little larger than the ideal learning scenario depicted in blue on the left, where the learning process uses the true stationary probabilities in the gradient expression, i.e. only the Q-differences are estimated for the computation of the gradient. The FVRL algorithm is closer than MC to such scenario.

We also observe that the non-trivial θ_1 parameter goes through large variations in all learning scenarios –see Figure 5(a)– which tend to happen when the value of θ_1 goes below its optimum of 3. This is due to the fact that the non-deterministic probability at $K_1 - 1$ of the acceptance policy suddenly changes from being close to 0 to being close to 1, which makes the learning agent suddenly receive a signal that it should reject incoming jobs at a larger threshold value.

[7] When the stationary distribution of the states in $\vec{\partial}\mathcal{A}^c$ is unknown, one should use the same procedure described in Subsection 4.2.1 for the selection of the start state.



(a) Learning paths of 20 replications



(b) Distribution of learning paths of 20 replications

Figure 5: Comparison of FVRL learning (middle) with Monte-Carlo learning (right) and learning using the true stationary probabilities (left), which gives the scenario closest to the true learning process, where only the difference in the Q values is estimated. The plots show the results over 20 experiments run over 30 learning steps on a loss network with $R = 6$ servers, $\lambda = [1, 5]$, $\rho = [0.5, 0.3]$, with blocking costs $C = [2 \times 10^3, 2 \times 10^4]$. At each learning step, the FVRL setup sets the absorption set size by job class at $J_i/K_i = [0.3, 0.5]$, the number of particles at $N = 300$ and the number of arrival steps at $T = 500$. In the MC learning case, each experiment uses as many number of events per learning step as the average number of events observed in FVRL by learning step. In all cases the initial guess is $\theta = [0.1, 0.1]$ ($K = [2, 2]$), the learning parameter starts at $\alpha = 1.0$ and is then decreased inversely proportional to the learning step. The top row shows each of the 20 realized paths. The bottom row shows the distribution of the visited θ values by learning step, where the darker band includes 50% of the paths around their median and the lighter band includes 100% of the paths. The interesting learning paths to compare are those associated to θ_1, as θ_2 is learned equally well as 4.9 by all methods (which is shown as dotted lines in (a)).

5 Conclusions and future work

The Fleming-Viot particle system was presented as an efficient alternative to Monte-Carlo for the exploration of environments where rewards are sparse and their occurrence is rare. Its application to the estimation of the blocking probability in an $M/M/1/K$ queueing system and of the expected rejection cost in a loss network serving different job classes served as test benches. In the $M/M/1/K$ system, the method proved to be much more efficient than Monte-Carlo for large capacities K, where the latter completely fails, and in the loss network system it was able to estimate the expected cost more accurately and more precisely than Monte-Carlo.

Results on optimal control of the above systems using pol-

icy gradient were also presented. The proposed FVRL algorithm tends to find the optimum parameters significantly faster than Monte-Carlo learning. In this case, we note that the accuracy of the stationary probability estimator is not as crucial as in the estimation problem, because the algorithm is able to learn as long as it receives a signal from the rare states.

In future work, we intend to extend the FV and the FVRL algorithms to environments other than queues and networks, to more traditional RL environments such as labyrinths or the mountain car, where it will be crucial to define the absorption set $\mathcal{A}$ adaptively, i.e. based on the discovery of the states that give no rewards during exploration of the environment.

Acknowledgements

This work was partially supported by the French National Research Agency through projects ANR-11-LABX-0040 of the "Investments for the Future" program and ANR-22-CE25-0013-02 (EPLER), and by project LAGOON of the Stic-amsud program.

References

[1] S. Asmussen. Applied Probability and Queues. Applications of mathematics : stochastic modelling and applied probability. Springer, 2003.

[2] A. Asselah, P.A. Ferrari, and P. Groisman. Quasistationary distributions and Fleming-Viot processes in finite spaces. J. Appl. Probab., 48(2):322–332, 2011.

[3] Thomas Bonald, Matthieu Jonckheere, and Alexandre Proutière. Insensitive load balancing. ACM Sigmetrics Performance Evaluation Review, 32(1):367–377, 2004.

[4] Léon Bottou, Frank E. Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. SIAM Review, 60(2):223–311, jan 2018.

[5] Pierre Brémaud. Markov chains: Gibbs fields, Monte Carlo simulation, and queues, volume 31. Springer Science & Business Media, 2013.

[6] Yuri Burda, Harri Edwards, Deepak Pathak, Amos Storkey, Trevor Darrell, and Alexei A Efros. Large-scale study of curiosity-driven learning. arXiv preprint arXiv:1808.04355, 2018.

[7] K. Burdzy, R. Holyst, and P. March. A Fleming-Viot particle representation of the Dirichlet Laplacian. Comm. Math. Phys., 214(3):679–703, 2000.

[8] Bertrand Cloez and Josué Corujo. Uniform in time propagation of chaos for a moran model. arXiv preprint arXiv:2107.10794, 2021.

[9] Bertrand Cloez and Marie-Noémie Thai. Quantitative results for the fleming–viot particle system and quasi-stationary distributions in discrete space. Stochastic Processes and their Applications, 126(3):680–702, 2016.

[10] J. N. Darroch and E. Seneta. On quasi-stationary distributions in absorbing continuous-time finite Markov chains. J. Appl. Probability, 4:192–196, 1967.

[11] Gabriel Dulac-Arnold, Daniel J. Mankowitz, and Todd Hester. Challenges of real-world reinforcement learning. CoRR, abs/1904.12901, 2019.

[12] P. Groisman and M. Jonckheere. Simulation of quasi-stationary distributions on countable spaces. Markov Process. Related Fields, 19(3):521–542, 2013.

[13] G. Yin H. J. Kushner. Stochastic Approximation and Recursive Algorithms and Applications. Springer-Verlag, 2003.

[14] Ger Koole. Structural results for the control of queueing systems using event-based dynamic programming. Queueing systems, 30(3):323–339, 1998.

[15] Ger Koole. Monotonicity in Markov reward and decision chains: Theory and applications, volume 1. Now Publishers Inc, 2007.

[16] Antonio Massaro, Francesco De Pellegrini, and Lorenzo Maggi. Optimal trunk-reservation by policy learning. In IEEE INFOCOM 2019, apr 2019.

[17] Maja J Mataric. Reward functions for accelerated learning. In Machine learning proceedings 1994, pages 181–189. Elsevier, 1994.

[18] Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, volume 70 of Proceedings of Machine Learning Research, pages 2778–2787. PMLR, 06–11 Aug 2017.

[19] Alexey Piunovskiy and Yi Zhang. Continuous-time markov decision processes. Probability Theory and Stochastic Modelling, 2020.

[20] M. L. Puterman. Markov Decision Processes: Discrete Stochastic Dynamic Programming. John Wiley & Sons, 2005.

[21] Keith W. Ross. Multiservice Loss Models for Broadband Telecommunication Networks. Springer, 1995.

[22] Keith W. Ross and Danny H.K. Tsang. The stochastic knapsack problem. IEEE Transactions on Communications, 37(7):740–747, July 1989.

[23] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. nature, 550(7676):354–359, 2017.

[24] Richard S Sutton, David A McAllester, Satinder P Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. In Advances in neural information processing systems, pages 1057–1063, 2000.

[25] Martin J Wainwright. High-dimensional statistics: A non-asymptotic viewpoint, volume 48. Cambridge University Press, 2019.

A Estimation of $\mathbb{E}^\pi(\eta)$ using Fleming-Viot particle systems

An estimator of $\mathbb{E}^\pi(\eta)$ using (3) is constructed from estimators $\hat{f}_\eta, \hat{g}$ of functions f_η, g as

$$\hat{\mathbb{E}}^\pi(\eta) = \int_0^\infty \hat{f}_\eta(t)\hat{g}(t)dt. \tag{12}$$

Firstly, we explain the construction of the estimator of g, which is a ratio of the quantities $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_\mathcal{K} > t)$ and $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_\mathcal{A})$, each of which is estimated from two separate simulations, as follows: (i) a simulation of the Markov process starting at an arbitrary state $x \in \vec{\partial}\mathcal{A}$, and (ii) a simulation of a predefined number of N independent copies of the Markov process starting at a randomly selected state $x \in \vec{\partial}\mathcal{A}^c$ following the entrance state distribution into $\mathcal{A}^c$ under stationarity, estimated from the first simulation. As explained below, simulations (i) and (ii) contribute to the estimation of $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_\mathcal{A})$ while simulation (ii) contributes to the estimation of $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_\mathcal{K} > t)$.

The details are as follows: we set $\tau_{\mathcal{A},0} \doteq 0$ and define a sequence of stopping times $\tau_{\mathcal{A}^c,i}$, $\tau_{\mathcal{A},i}$, associated to the events of entry, in the i-th cycle, into $\mathcal{A}^c$ and into $\mathcal{A}$, respectively, as follows:

$$\begin{aligned}\tau_{\mathcal{A}^c,i} &= \inf_{t>\tau_{\mathcal{A},i-1}}\{X_t^\pi \in \mathcal{A}^c\},\\ \tau_{\mathcal{A},i} &= \inf_{t>\tau_{\mathcal{A}^c,i}}\{X_t^\pi \in \mathcal{A}\},\end{aligned}$$

for $i \geq 1$. We also define $T_{E,i} = \tau_{\mathcal{A}^c,i} - \tau_{\mathcal{A},i-1}$, the entry time into $\mathcal{A}^c$ within cycle i, and $T_{\mathcal{K},i} = \tau_{\mathcal{A},i} - \tau_{\mathcal{A}^c,i}$, the killing time within cycle i. We note that $T_{\mathcal{A},i} = T_{E,i} + T_{\mathcal{K},i}$, which will be used below to construct the estimator of the denominator of g, $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_\mathcal{A})$.

The first simulation consists of running the process X_t^π until a predefined number $M_0 + M$ entry times $\{T_{E,i}\}_{i=1}^{M_0+M}$ are observed, where we consider the first M_0 observations to be burn-in in order to assume stationarity of the process thereafter. From this simulation we compute the empirical entrance state distribution into $\mathcal{A}^c$, which can be considered as an estimator of $p^\pi_{\vec{\partial}\mathcal{A}^c}$, as $\hat{p}^\pi_{\vec{\partial}\mathcal{A}^c} = \frac{1}{M}\sum_{i=M_0+1}^{M_0+M} \mathbf{1}_{X^\pi_{\tau_{\mathcal{A}^c,i}}}$.

The second simulation of the N independent copies of X_t^π is started at states $x_i \in \vec{\partial}\mathcal{A}^c$ randomly chosen according to the estimated stationary distribution $\hat{p}^\pi_{\vec{\partial}\mathcal{A}^c}$ (this is trivial when $\vec{\partial}\mathcal{A}^c$ has a single state), yielding N killing times $\{T_{\mathcal{K},i}\}_{i=1}^N$.

From the last M entry times and the N killing times, we define a Monte-Carlo estimator of g as the ratio

$$\hat{g}(t) = \frac{\frac{1}{N}\sum_{i=1}^N \mathbf{1}_{T_{\mathcal{K},i}>t}}{\frac{1}{M}\sum_{i=M_0+1}^{M_0+M} T_{E,i} + \frac{1}{N}\sum_{i=1}^N T_{\mathcal{K},i}}. \tag{13}$$

Next, we explain the construction of the estimator of f, which is a η-weighted sum of $\phi_t^{\vec{\partial}\mathcal{A}^c}$ over all states in $\mathcal{A}^c$, using the Fleming-Viot N-particle system driven by X_t^π. The Fleming-Viot system, denoted by $(\xi_t^\nu)_{t\geq 0}$, is simulated as described in Section 3.2, using $\mathcal{A}$ as the absorption set and starting each particle i at x_i a randomly chosen state according to the distribution $\nu(x) = p^\pi_{\vec{\partial}\mathcal{A}^c}(x)\mathbf{1}_{x\in\vec{\partial}\mathcal{A}^c}$. That is, all particles start at the boundary of $\mathcal{A}^c$, as required by the quantity to estimate $\phi_t^{\vec{\partial}\mathcal{A}^c}$. We let $m(\cdot,\xi) : \mathcal{A}^c \to [0,1]$ denote the empirical distribution of the N particles with positions described by vector ξ, defined as the empirical mean $m(x,\xi) \doteq \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{\xi(i)=x}, \forall x \in \mathcal{A}^c$. Since $m(\cdot,\xi_t^\nu)$ is an estimator of $\phi_t^{\vec{\partial}\mathcal{A}^c}$ (because ν is restricted to the boundary of $\mathcal{A}^c$ where it is equal to the entrance state distribution into $\mathcal{A}^c$ under stationarity), f can be estimated by

$$\hat{f}_\eta(t) = \sum_{x\in\mathcal{A}^c} \eta(x)m(x,\xi_t^\nu). \tag{14}$$

We note that, by construction, $\hat{g}(t) = 0$ for $t > T_{\mathcal{K},\max} = \max\{T_{\mathcal{K},i} : 1 \leq i \leq N\}$. Therefore, since we wish to compute $\hat{\mathbb{E}}^\pi(\eta) = \int_0^\infty \hat{f}_\eta(t)\hat{g}(t)dt$, we only need to simulate the Fleming-Viot process until time $T_{\mathcal{K},\max}$ is reached. Also, since both $\hat{f}_\eta(t)$ and $\hat{g}(t)$ are almost surely piecewise constant functions and $\hat{g}(t) = 0$ for $t > T_{\mathcal{K},\max}$, the integral $\int_0^\infty \hat{f}_\eta(t)\hat{g}(t)dt$ is a finite sum that can be easily computed.

B Proof of Theorem 2: Error of the Fleming-Viot estimator

Using the notation $f_\eta, g, \hat{f}_\eta, \hat{g}$ introduced in (3) and (14), (13), we have $\mathbb{E}^\pi(\eta) = \int_0^\infty f_\eta(t)g(t)dt$ and $\hat{\mathbb{E}}^\pi(\eta) = \int_0^\infty \hat{f}_\eta(t)\hat{g}(t)dt$. We are thus interested in bounding:

$$\mathbb{E}\left|\int_0^\infty \hat{f}_\eta \hat{g} dt - \int_0^\infty f_\eta g dt\right|$$

We start by decomposing the problem of upper bounding the above quantity into two subproblems in the following way:

$$\mathbb{E}\left|\int_0^\infty \hat{f}_\eta \hat{g} dt - \int_0^\infty f_\eta g dt\right| \leq \mathbb{E}\left|\int_0^\infty (f_\eta - \hat{f}_\eta) g dt\right| + \mathbb{E}\left|\int_0^\infty \hat{f}_\eta(\hat{g} - g) dt\right| \tag{15}$$

We start by bounding the first term on the right hand side. For this purpose we will need the uniform propagation of chaos bound presented in (4), that is:

$$\sup_{\|\phi\|_\infty \leq 1} \sup_{t\geq 0} \mathbb{E}\Big|[m(\cdot, \xi_t^\nu)(\phi)] - \phi_t^\nu(\phi)\Big| \leq \frac{C_{\text{FV}}}{\sqrt{N}},$$

from which it follows that:

$$\sup_{t\geq 0} \mathbb{E}\Big|[\hat{f}_\eta(t) - f_\eta(t)\Big| \leq \frac{C_{\text{FV}}}{\sqrt{N}}. \tag{16}$$

As was mentioned in subsection 3.3, this bound follows directly from [8, Theorem 1.4]. The assumptions of [8, Theorem 1.4] have a very general form, but it is easy to check that they are trivially satisfied in our simple case. The assumption (I) on initialization is satisfied by our assumption that the FV particle system is started at the position of N i.i.d samples from $p^\pi_{\vec{\delta}_{\mathcal{A}^c}}$. Assumption (C1) has several parts: the uniform bound on selection rates (which are, in our case, the rates of jumps out of $\mathcal{A}^c$), follows from the fact that the state space is finite); the rest of assumption (C1) is trivially satisfied when we take $V_\mu^d(x)$ to be the rate of jump out of $\mathcal{A}^c$ from the state $x \in \mathcal{A}^c$ for any μ, and set function $V_\mu^b(y), V_\mu^s(x,y)$ equal to zero. Finally, assumption (C2) follows from the fact that we are working with an irreducible Markov chain on a finite state space. Therefore, using the triangle inequality and the inequality (16), we obtain:

$$\begin{aligned}
\mathbb{E}\left|\int_0^\infty (f_\eta - \hat{f}_\eta) g dt\right| &\leq \mathbb{E}\int_0^\infty \left|(f_\eta - \hat{f}_\eta)\right| g dt \\
&\leq \int_0^\infty \mathbb{E}\left|(f_\eta - \hat{f}_\eta)\right| g dt \\
&\leq \frac{C_{\text{FV}}}{\sqrt{N}} \int_0^\infty g dt = \frac{\mathbb{E}_{\vec{\delta}_{\mathcal{A}^c}} T_{\mathcal{K}}}{\mathbb{E}_{\vec{\delta}_{\mathcal{A}}}(T_{\mathcal{A}})} \frac{C_{\text{FV}}}{\sqrt{N}},
\end{aligned}$$

where in the last line we also use the 'wedding cake decomposition', $\int_0^\infty \mathbb{P}_{\vec{\delta}_{\mathcal{A}^c}}(T_{\mathcal{K}} > t) dt = \mathbb{E}_{\vec{\delta}_{\mathcal{A}^c}} T_{\mathcal{K}}$.

Since $\eta(x)$ is bounded, without loss of generality we can consider that $\sup_{x\in\mathcal{A}^c} |\eta(x)| \leq 1$, and use that $\left|\hat{f}_\eta(t)\right| \leq 1$ for $t \geq 0$ to get:

$$\left|\int_0^\infty \hat{f}_\eta(\hat{g} - g) dt\right| \leq \int_0^\infty |\hat{g} - g|\, dt.$$

We thus wish to estimate $\mathbb{E}\int_0^\infty |\hat{g} - g|\, dt$. For convenience, we define a random variable $\overline{T}_{\mathcal{A}}$ with the distribution of $T_{\mathcal{A}}$ when X_t^π is started with distribution $p^\pi_{\vec{\delta}_{\mathcal{A}}}$, and a random variable $\overline{T}_{\mathcal{K}}$ with the distribution of $T_{\mathcal{K}}$ when X_t^π is started with distribution $p^\pi_{\vec{\delta}_{\mathcal{A}^c}}$. Since we start the simulation of X_t^π for the purpose of estimating g with distribution $p^\pi_{\vec{\delta}_{\mathcal{A}}}$, we do not need any burn-in. We therefore take $M_0 = 0$. We also note that since we start the simulation at the distribution $p^\pi_{\vec{\delta}_{\mathcal{A}}}$, it follows from renewal theory [1] that the inter-arrival times $\mathcal{T}_{\mathcal{A},i}$ used to construct the estimator $\hat{g}$ are i.i.d. with distribution $\overline{T}_{\mathcal{A}}$.

We also introduce additional shorthand notation for the numerators and denominators of $g(t)$ and $\hat{g}(t)$. We denote $N_t = \mathbb{P}_{\vec{\delta}_{\mathcal{A}^c}}(T_{\mathcal{K}} > t)$ and $D_{\mathcal{A}} = \mathbb{E}_{\vec{\delta}_{\mathcal{A}}} T_{\mathcal{A}}$, $D_{\mathcal{K}} = \mathbb{E}_{\vec{\delta}_{\mathcal{A}^c}} T_{\mathcal{K}}$. We also denote by $\hat{N}_t, \hat{D}_{\mathcal{A}}$ the estimators of $N_t, D_{\mathcal{A}}$, that is $\hat{N}_t = \frac{1}{N}\sum_{i=1}^N \mathbf{1}_{T_{\mathcal{K},i} > t}$ and $\hat{D}_{\mathcal{A}} = \hat{D}_{\mathbb{E}} + \hat{D}_{\mathcal{K}}$, with

$$\hat{D}_{\mathbb{E}} = \frac{1}{M}\sum_{i=1}^M T_{E,i},$$

$$\hat{D}_{\mathcal{K}} = \frac{1}{N}\sum_{i=1}^N T_{\mathcal{K},i},$$

We thus have $g(t) = \frac{N_t}{D}$ and $\hat{g}_t = \frac{\hat{N}_t}{\hat{D}_{\mathcal{A}}}$.

We are interested in bounding:

$$\mathbb{E}\int_0^\infty \left|\frac{\hat{N}_t}{\hat{D}_{\mathcal{A}}} - \frac{N_t}{D_{\mathcal{A}}}\right|$$

Using the triangle inequality $\left|\frac{\hat{N}_t}{\hat{D}_\mathcal{A}} - \frac{N_t}{D_\mathcal{A}}\right| \leq \left|\frac{\hat{N}_t}{\hat{D}_\mathcal{A}} - \frac{\hat{N}_t}{D_\mathcal{A}}\right| + \left|\frac{\hat{N}_t}{D_\mathcal{A}} - \frac{N_t}{D_\mathcal{A}}\right|$ we get:

$$\mathbb{E}\int_0^\infty \left|\frac{\hat{N}_t}{\hat{D}_\mathcal{A}} - \frac{N_t}{D_\mathcal{A}}\right| dt \leq \mathbb{E}\int_0^\infty \hat{N}_t \left|\frac{1}{\hat{D}_\mathcal{A}} - \frac{1}{D_\mathcal{A}}\right| dt + \mathbb{E}\int_0^\infty \frac{1}{D_\mathcal{A}} \left|\hat{N}_t - N_t\right| dt,$$

Using the formula $\int_0^\infty \hat{N}_t dt = \frac{1}{N}\sum_{i=1}^N T_{\mathcal{K},i} = \hat{D}_\mathcal{K}$, the first term on the right hand side of the above bound is equal to $\mathbb{E}\left|\frac{\hat{D}_\mathcal{K}}{\hat{D}_\mathcal{A}} - \frac{\hat{D}_\mathcal{K}}{D_\mathcal{A}}\right|$. To bound this quantity, we introduce an event $B = \{\hat{D}_\mathcal{A} < \frac{1}{2}D_\mathcal{A}\}$. We use the decomposition

$$\mathbb{E}\left|\frac{\hat{D}_\mathcal{K}}{\hat{D}_\mathcal{A}} - \frac{\hat{D}_\mathcal{K}}{D_\mathcal{A}}\right| = \mathbb{E}\mathbf{1}_B \left|\frac{\hat{D}_\mathcal{K}}{\hat{D}_\mathcal{A}} - \frac{\hat{D}_\mathcal{K}}{D_\mathcal{A}}\right| + \mathbb{E}\mathbf{1}_{B^c} \left|\frac{\hat{D}_\mathcal{K}}{\hat{D}_\mathcal{A}} - \frac{\hat{D}_\mathcal{K}}{D_\mathcal{A}}\right| \tag{17}$$

and bound each of the terms separately.

Since we always have $0 \leq \frac{\hat{D}_\mathcal{K}}{\hat{D}_\mathcal{A}} \leq 1$ and on the set B we have $0 \leq \frac{\hat{D}_\mathcal{K}}{D_\mathcal{A}} \leq \frac{1}{2}$, we have:

$$\mathbb{E}\mathbf{1}_B \left|\frac{\hat{D}_\mathcal{K}}{\hat{D}_\mathcal{A}} - \frac{\hat{D}_\mathcal{K}}{D_\mathcal{A}}\right| \leq \mathbb{P}(B).$$

Since the Markov Process X_t^π is irreducible and the state space $\mathcal{S}$ is finite, it is geometrically ergodic [5]. It follows then that there exist constants $C, \lambda > 0$, such that $\mathbb{P}(\overline{T}_\mathcal{A} > t) \leq C\exp(-\lambda t)$. Therefore, by [25, Theorem 2.13], the random variable $\overline{T}_\mathcal{A}$ is subexponential. Furthermore $\mathbb{P}(B) \leq \mathbb{P}_{\vec{\partial}\mathcal{A}^c}\left(\left|\hat{D}_\mathcal{A} - D_\mathcal{A}\right| \geq \frac{1}{2}D_\mathcal{A}\right)$. From the concentration bound for the standard estimator of the mean of subexponential variables [25][Equation 2.18], it follows that there exists $c > 0$:

$$\mathbb{P}(B) \leq e^{-c\sqrt{1/N + 1/M}}.$$

To bound the second term in (17), we observe that the function $h(x) = 1/x$ is Lipschitz continuous on $[a, \infty)$ for any $a > 0$, with the Lipschitz constant $L_a = \sup_{x\in[a,\infty)} |h'(x)| = \frac{1}{a^2}$. Using this fact with $a = D_\mathcal{A}/2$, we have:

$$\mathbb{E}\mathbf{1}_{B^c}\hat{D}_\mathcal{K}\left|\frac{1}{\hat{D}_\mathcal{A}} - \frac{1}{D_\mathcal{A}}\right| \leq \frac{4}{D_\mathcal{A}^2}\mathbb{E}\hat{D}_\mathcal{K}\left|\hat{D}_\mathcal{A} - D_\mathcal{A}\right|.$$

Using the Cauchy-Schwartz inequality, we get:

$$\begin{aligned}
\mathbb{E}\hat{D}_\mathcal{K}\left|\hat{D}_\mathcal{A} - D\right| &\leq \left(\mathbb{E}\hat{D}_\mathcal{K}^2\right)^{1/2}\left(\mathbb{E}\left|\hat{D}_\mathcal{A} - D_\mathcal{A}\right|^2\right)^{1/2} \\
&\leq \left(\left(\mathbb{E}\hat{D}_\mathcal{K}\right)^2 + \mathrm{Var}(\hat{D}_\mathcal{K})\right)^{1/2}\left(\mathrm{Var}(\hat{D}_\mathcal{A})\right)^{1/2} \\
&\leq \Big[\frac{\left(\mathrm{Var}_{\vec{\partial}\mathcal{A}}T_E\right)^{1/2}}{\sqrt{M}} + \frac{\left(\mathrm{Var}_{\vec{\partial}\mathcal{A}}T_\mathcal{K}\right)^{1/2}}{\sqrt{N}}\Big]\left((D_\mathcal{K})^2 + \frac{1}{M}\mathrm{Var}_{\vec{\partial}\mathcal{A}^c}(D_\mathcal{K})\right)^{1/2}
\end{aligned}$$

We therefore obtain

$$\mathbb{E}\int_0^\infty \hat{N}_t\left|\frac{1}{\hat{D}_\mathcal{A}} - \frac{1}{D_\mathcal{A}}\right| dt \leq \mathbb{E}_{\vec{\partial}\mathcal{A}^c}T_\mathcal{K}\Big[\frac{\left(\mathrm{Var}_{\vec{\partial}\mathcal{A}}T_E\right)^{1/2}}{\sqrt{M}} + \frac{\left(\mathrm{Var}_{\vec{\partial}\mathcal{A}}T_\mathcal{K}\right)^{1/2}}{\sqrt{N}}\Big] + \mathcal{O}(\frac{1}{M}).$$

We are left with bounding

$$\frac{1}{D_\mathcal{A}}\mathbb{E}\int_0^\infty \left|\hat{N}_t - N_t\right| dt.$$

Since $\hat{N}_t$ is an average of M Bernoulli random variables with mean N_t, we have:

$$\begin{aligned}
\frac{1}{D_\mathcal{A}}\mathbb{E}\int_0^\infty \left|\hat{N}_t - N_t\right| dt &= \frac{1}{D_\mathcal{A}}\int_0^\infty \mathbb{E}\left|\hat{N}_t - N_t\right| dt \\
&\leq \frac{1}{D_\mathcal{A}}\int_0^\infty \sqrt{\mathbb{E}\left|\hat{N}_t - N_t\right|^2} dt \\
&= \frac{1}{\sqrt{N}D_\mathcal{A}}\mathbb{E}\int_0^\infty \sqrt{N_t(1-N_t)},
\end{aligned}$$

where in the first inequality we use $\mathbb{E}Y \leq \sqrt{\mathbb{E}Y^2}$ which follows from Cauchy-Schwartz inequality. We note, that $N_t = \mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_\mathcal{K} > t) = 1 - \mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_\mathcal{K} \leq t) = 1 - F_\mathcal{K}(t)$. Thus we have

$$\frac{1}{D_\mathcal{A}}\mathbb{E}\int_0^\infty \left|\hat{N}_t - N_t\right| dt \leq \frac{1}{\sqrt{N}\mathbb{E}_{\vec{\partial}\mathcal{A}}T_\mathcal{A}}\int_0^\infty \sqrt{F_\mathcal{K}(t)(1-F_\mathcal{K}(t))}dt.$$

Combining all of the above inequalities in an obvious manner, we obtain the bound from the thesis.

It follows from exponential ergodicity of X_t^π that $\overline{T}_\mathcal{A}, \overline{T}_\mathcal{K}$ have exponential tails, that is, there exist constants $C_\mathcal{A}, \lambda_\mathcal{A}, C_\mathcal{K}, \lambda_\mathcal{K}$ such that $\mathbb{P}_{\vec{\partial}\mathcal{A}}(T_\mathcal{A} > t) \leq C_\mathcal{A}\exp(-\lambda_\mathcal{A} t)$ and $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_\mathcal{K} > t) \leq C_\mathcal{K}\exp(-\lambda_K t)$. Therefore all the moments and the integral above are finite.

C Heuristics for the choice of the simulation parameters of the Fleming-Viot estimator

In this section we leverage theoretical results on the well-studied $M/M/1/K$ queue system to provide insights about the appropriate choice of parameter J –which in this unidimensional case fully defines the size of the absorption set $\mathcal{A}$– and adapting the values of the number of particles N and of the number of arrival events T (which directly impacts the number of cycles M defined in Appendix B controlling, together with N, the convergence rate of the FV estimator) with the goal of obtaining an accurate estimation of the blocking probability, i.e. of the stationary probability of state K. The conclusions of this analysis could be used as an initial guideline for the choice of these parameters in a more general setting such as multidimensional problems.

We focus the analysis on the two quantities (out of the three involved in expression (7)) whose estimation is highly affected by the trade-off of the choice of J described in remark 3, $\phi_t^J(K)$ and $\mathbb{E}_{J-1}(T_{\mathcal{A}})$, since the third quantity, $\mathbb{P}_J(T_{\mathcal{K}} > t)$, does not depend strongly on J (see Appendix B)[8]. The trade-off described in the remark states that, for constant values of N and T, the error in the estimation of $\phi_t^J(K)$ tends to decrease as J increases, while the error in the estimation of $\mathbb{E}_{J-1}(T_{\mathcal{A}})$ tends to increase. For the purpose of this impact analysis, we invert our reasoning: we fix J, and set the values of N and T required to approximately satisfy predefined expected relative errors in the respective estimators. We then run simulations using each N-T pair dictated by each relative error pair considered, and study statistics on the relative error obtained in the FV estimator of the blocking probability as a function of the values of N and T.

The values of N and T for each expected relative error pair are determined based on the following heuristics that approximately express the two expected relative errors as a function of J, N, T, and the system's characteristics:

1. **Relative error of $\hat{\phi}_t^J(K)$:** Considering that in an $M/M/1/K$ queue system with $\rho < 1$, the conditioned blocking probability $\phi_t^{J=1}(K)$ converges as $t \to \infty$ towards $\sqrt{K}\rho^{K/2}$ [10], i.e. $\sim O(\rho^{K/2})$, similarly, the conditioned blocking probability for any absorption set size J, $\phi_t^J(K)$, increases to $\sim O(\rho^{(K-J+1)/2})$ as $t \to \infty$. As a very rough approximation and to get order of magnitude relations, we could think of $\phi_t^J(K)$ as the blocking probability q of a single-server queue system with capacity equal to $\lceil \frac{(K-J+1)}{2} \rceil$, namely $q = \frac{(1-\rho)\rho^{K_J}}{1-\rho^{K_J+1}}$, where $K_J = \lceil (K-J+1)/2 \rceil$. Under this assumption, and ignoring the interdependence among FV particles, the relative error of estimating $\phi_t^J(K)$ with the empirical mean on N samples is derived from the variance of a Binomial(q) random variable as $\epsilon_\phi \sim \sqrt{(1-q)/Nq}$. Thus, given q, the value of N to approximately satisfy a desired expected relative error ϵ_ϕ for $\hat{\phi}_t^J(K)$ is obtained as $N \sim \lceil \frac{1-q}{q\epsilon_\phi^2} \rceil \approx \lceil \frac{1}{q\epsilon_\phi^2} \rceil$, if $q \ll 1$ which is usually the case.

2. **Relative error of $\hat{\mathbb{E}}_{J-1}(T_{\mathcal{A}})$:** In an $M/M/1/K$ queue system with $\rho < 1$, the expected return time to the state $J-1$, when starting at $J-1$, is equal to $\mathbb{E}_{J-1}(T_{J-1}) = \frac{1}{\lambda p(1+\rho^{-1})}$ [5], where p is the stationary probability of the state $x = J-1$, i.e. $p = \frac{(1-\rho)\rho^{J-1}}{1-\rho^{K+1}}$. Given the absorption set $\mathcal{A} = \{0, 1, ..., J-1\}$, it can easily be shown that the expected return time to $\mathcal{A}$, when starting at $J-1$, is equal to $\mathbb{E}_{J-1}(T_{\mathcal{A}}) = \mathbb{E}_{J-1}(T_{J-1})(1+\rho^{-1})$. If we want to observe M return cycles to $\mathcal{A}$, we should simulate the queue system for as long as $t_S = M\mathbb{E}_{J-1}(T_{\mathcal{A}}) = M\mathbb{E}_{J-1}(T_{J-1})(1+\rho^{-1})$, i.e. $t_S = \frac{M}{\lambda p}$. Since λt_S is the expected number of arrival events T observed in the time span t_S, we should simulate the system for as long as $T \sim M/p$ arrival events. On the other hand, given M return cycles to $\mathcal{A}$, the standard error of the moment estimator of $\mathbb{E}_{J-1}(T_{\mathcal{A}})$ is given by $\sigma(T_A)/\sqrt{M}$. It is reasonable to assume (confirmed by experiments) that $\sigma(T_A) \sim \mathbb{E}_{J-1}(T_A)$, hence the relative error of the estimator of $\mathbb{E}_{J-1}(T_A)$, ϵ_{ET}, is of the order of $1/\sqrt{M}$, which makes $M \sim 1/\epsilon_{ET}^2$. Thus, given p, the value of T to approximately satisfy a desired expected relative error ϵ_{ET} for $\hat{\mathbb{E}}_{J-1}(T_{\mathcal{A}})$ is obtained as $T \sim \lceil \frac{1}{p\epsilon_{ET}^2} \rceil$.

From the above, the following common aspects are observed about the minimum values of N and T required to satisfy predefined expected relative errors in the estimators of $\phi_t^J(K)$ and $\mathbb{E}_{J-1}(T_{\mathcal{A}})$:

1. N affects the relative error of $\hat{\phi}_t^J(K)$ and T affects the relative error of $\hat{\mathbb{E}}_{J-1}(T_{\mathcal{A}})$.
2. The relationship of N and T with their respective relative errors is of the same form, i.e. proportional to the inverse of a stationary probability and to the inverse of the squared expected relative error.

On the other hand, the following difference is observed: for $\rho < 1$ and sufficiently large K[9], the value of N as a function of the stationary probability q is dominated by an increasing exponential function of $(K-J)/2$, i.e. $\sim O(\rho^{-(K-J)/2})$ whereas the value of T as a function of the stationary probability p is dominated by an increasing exponential function of J, i.e. $\sim O(\rho^{-J})$ which, importantly, does not depend on K. Thus, as mentioned in remark 3, the closer is J to 0, the smaller the required T and the larger the required N for fixed expected relative errors, while the opposite is true when J gets closer to K.

[8] Although the error of the estimator of $\mathbb{P}_J(T_{\mathcal{K}} > t)$ depends on N, it is sensible to assume that its estimation error decreases as N increases similarly or faster than the error of the estimator of $\phi_t^J(K)$, so we can limit our analysis to the error of the estimator of the latter.

[9] K is considered sufficiently large in this context when ρ^{K-J} can be neglected w.r.t. 1.

More importantly, experiments showed that the computational complexity –in terms of number of observed events– required to satisfy a given ϵ_ϕ value (which impacts the number of observed events of the FV simulation) is much larger than the computational complexity required to satisfy the same value in ϵ_{ET} (which impacts the number of observed events in the single simulation of the X_t^π process). Concrete values of the respective computational complexities can be derived from the results shown in Figure 2 and are complemented by the related observations presented within that Section 4.1.1. Thus, from the computational perspective alone, it is more convenient to favour a smaller error in $\hat{\mathbb{E}}_{J-1}(T_{\mathcal{A}})$ than a smaller error in $\hat{\phi}_t^J(K)$. Below we will see that this is also the case in terms of the estimation error of the FV estimator.

We completed the study of the appropriate choice of N and T by analyzing the impact of the errors of estimating $\phi_t^J(K)$ and $\mathbb{E}_{J-1}(T_{\mathcal{A}})$ on the FV estimator of the blocking probability. To this end, we ran experiments on different combinations of expected relative errors ϵ_ϕ and ϵ_{ET} and measured the accuracy in the estimation of the blocking probability of an $M/M/1/K$ system with $\rho = \lambda = 0.7$, $K = 20$, using a constant absorption set size of $J = 12$. The results of these experiments are shown in Figure 6 in terms of the estimation accuracy of the blocking probability $\hat{p}_{FV}(K)$.

From this heatmap, we conclude that it is more important to control the relative error in $\hat{\mathbb{E}}_{J-1}(T_{\mathcal{A}})$ than the relative error in $\hat{\phi}_t^J(K)$, as the contour lines are almost parallel to the axis where the relative error in $\hat{\phi}_t^J(K)$ is plotted, and their values indicate that the estimated blocking probability is larger than 1.5 times the true blocking probability when the relative error in $\hat{\phi}_t^J(K)$ is larger than $30\% - 40\%$. Note that, for each experiment run, the FV estimator of the blocking probability was computed only when a minimum of 5 return cycles to $J-1$ were observed (these cycles are used to estimate the denominator $\mathbb{E}_{J-1}(T_{\mathcal{A}})$) after the 10 initial transitions of the system. This 10 transitions were used as a burn-in period to allow the system to get closer to the stationary regime. Therefore, if the number of arrival events T is not large enough, the sample size on which the plotted median FV estimator is computed may be smaller than the 7 replications used for each N-T combination.

In Figure 2 we used these heuristics to choose the different values of N and T on which the convergence properties of the FV estimator were analyzed: given $J = 12$, for the left plots (a) and (c), T was fixed at the value associated to an approximate expected relative error in $\hat{\mathbb{E}}_{J-1}(T_{\mathcal{A}})$ equal to $\epsilon_{ET} = 20\%$, whereas the values of N were chosen for approximate expected relative errors in $\hat{\phi}_t^J(K)$ equal to $\epsilon_\phi = 20\%, 10\%, 5\%$ for $K = 20$, and equal to $\epsilon_\phi = 80\%, 60\%, 40\%$[10]; for the right plots (b) and (d), N was fixed at the value associated to an approximate expected relative error in $\hat{\phi}_t^J(K)$ equal to $\epsilon_\phi = 60\%$, whereas the values of T were chosen for approximate expected relative errors in $\hat{\mathbb{E}}_{J-1}(T_{\mathcal{A}})$ equal to $\epsilon_{ET} = 40\%, 20\%$[11].

[10] The larger relative errors chosen for $K = 40$ compared to $K = 20$ have to do with obtaining the same orders of magnitude for N in each K scenario.

[11] The expected relative error ϵ_{ET} depends on K only through $1 - \rho^{K+1}$, and this term can be safely approximated by 1 for large enough values of K such as 20 and 40 when $\rho = 0.7$. Thus, doing this approximation, the value of T satisfying a given ϵ_{ET} only depends on J and is thus the same for both K scenarios.

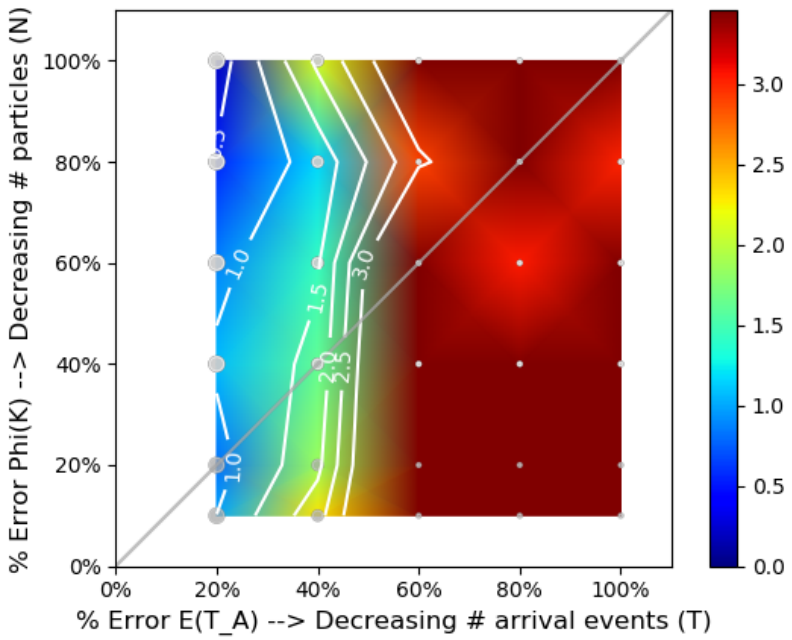


Figure 6: Heatmap showing the impact of different combinations of the expected relative errors for $\hat{\phi}_t^J(K)$ and $\hat{\mathbb{E}}_{J-1}(T_A)$ (indicated respectively on the vertical and horizontal axis) on the accuracy of the FV estimator of the blocking probability, $\hat{p}_{FV}(K)$, in an $M/M/1/K$ queue system over up to 7 valid experiments carried out for each combination. The system characteristics are $K = 20, \lambda = 0.7, \mu = 1$. The accuracy is shown as the ratio between the median $\hat{p}_{FV}(K)$ value and the true blocking probability $p(K)$. Thus, a ratio of 1 (light blue) implies 100% median accuracy, a ratio smaller than 1 (dark blue) implies underestimation, and a ratio larger than 1 (from cyan to red) implies overestimation. Selected contour levels are overlaid. The gray diagonal represents the line of equal expected relative errors in the two analyzed dimensions, and the gray points indicate the $5 \times 6 = 30$ error combinations on which experiments were run, whose size is proportional to the number of experiments (7 at the largest points, down to 2 at the smallest points). Smaller points are associated to a larger expected relative error in the estimation of $\mathbb{E}_{J-1}(T_{\mathcal{A}})$ which tend to preclude its estimation due to an insufficient number of observed return cycles to the absorption set $\mathcal{A}$, as described in the text. The color scale is chosen as the transformation $log_2(1 + z)$ where z is the ratio between the estimated and the true blocking probability; note that this transformation maps 0 to 0 and 1 to 1.

D FV and FVRL algorithms

This section presents the two algorithms used throughout this work to apply the Fleming-Viot methodology to network systems in order to (i) estimate the expected rejection cost (FV algorithm) and (ii) learn the optimum blocking sizes (FVRL algorithm).

FV algorithm

Data:

A. System characteristics: a loss network with R servers serving jobs of I different classes whose state is represented by the number of jobs of each class in the system, $\mathbf{x} = (x_1, ..., x_I)$.

B. System dynamics: the job arrival rate λ_i and the service rate μ_i for each job class $i = 1, \dots, I$ are known.

C. Job acceptance policy: an incoming job of class i is either accepted ($a = 1$) or rejected ($a = 0$). It is accepted whenever the system is not operating at full capacity R and when the number of jobs of the arriving class being served by the system is less than a constant K_i, $i = 1, \dots, I$. Otherwise, it is rejected, in which case a cost C_i is accrued. Thus, the job acceptance policy, when the system is at state $\mathbf{x}$ and a job of class i arrives ($\mathcal{L}_i$), is $\pi(a = 1|\mathbf{x}, \mathcal{L}_i) = \mathbf{1}_{\{\sum_{j=1}^{I} x_j < R\}} \mathbf{1}_{\{x_i < K_i\}}$, and the set of blocking states is $\mathcal{C} = \{\mathbf{x} : \sum_{j=1}^{I} x_j = R \text{ or } \sum_{j=1}^{I} x_j < R, x_i = K_i \text{ for some } i = 1, \dots I\}$.

D. $J_i, i = 1, \dots, I, 0 \leq J_i \leq K_i$: size of the absorption set $\mathcal{A}$ in dimension i, that is whenever the state of an FV particle visits a state $\mathbf{x}$ having $x_i = J_i - 1$ for some i, the particle is considered absorbed.

E. N: number of FV particles used to estimate $\phi_t^{\vec{\partial}\mathcal{A}^c}(K)$ and $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)$ in (3).

F. T: number of arrival events (incoming jobs), which has a direct impact on the number of cycles M used to estimate $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_{\mathcal{A}})$ in (3).

G. B: number of burn-in time steps to assure stationarity, typically 10-20.

H. M_0: minimum number of return cycles to $\mathcal{A}$ to have a reliable estimate of $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_{\mathcal{A}})$, typically 5-10.

Result: An estimate of the expected rejection cost under stationarity.

Steps: The algorithm is divided into the following three steps:

1. Estimation of $\mathbb{E}_{\vec{\partial}\mathcal{A}}(T_{\mathcal{A}})$:
   (a) Simulate the continuous-time Markov process $\mathbf{X}_t^{\pi}$ starting at $\mathbf{X}_0^{\pi}$ chosen uniformly at random from the set $\vec{\partial}\mathcal{A} = \{\mathbf{x} : x_i = J_i - 1 \text{ for some } i = 1, \dots, I\}$.
   (b) Record the observed values of the cycle time $T_{\mathcal{A}}$ defined in Section 3.1, measured as return times to the set $\mathcal{A}$ after the first visit to a state in $\vec{\partial}\mathcal{A}$ following the burn-in period given by parameter B.
   (c) Record the states at which the process enters $\mathcal{A}^c$.
   (d) Stop the simulation when the number of arrival events is equal to T, and record the number M of observed complete return cycles to $\mathcal{A}$ after the burn-in period.
   (e) If $M > M_0$, estimate $\mathbb{E}^{\vec{\partial}\mathcal{A}}(T_{\mathcal{A}})$ as $\frac{1}{M}\sum_{k=1}^{M} T_{\mathcal{A},k}$, o.w. conclude that the expected rejection cost cannot be reliably estimated, and end the process here.
   (f) Estimate the entrance distribution to $\mathcal{A}^c$, $\hat{p}^{\pi}_{\vec{\partial}\mathcal{A}^c}(\mathbf{x})$ as $\frac{1}{\overline{M}}\sum_{k=1}^{\overline{M}} \mathbf{1}_{\mathbf{x}(k)=\mathbf{x}}, \forall \mathbf{x} \in \vec{\partial}\mathcal{A}^c$, where $\overline{M}$ is the number of observed entrances to $\mathcal{A}^c$ at states $\mathbf{x}(k)$.
2. Estimation of $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)$ and $\phi_t^{\vec{\partial}\mathcal{A}^c}(\mathbf{x}_{\mathcal{C}})$, where $\mathbf{x}_{\mathcal{C}}$ is any state in the set of blocking states $\mathcal{C}$:
   (a) Simultaneously simulate N trajectories (FV particles) that independently follow the law of the Markov process $\mathbf{X}_t^{\pi}$, starting at $\mathbf{X}_0^{\pi}$ chosen in $\vec{\partial}\mathcal{A}^c$ following $\hat{p}^{\pi}_{\vec{\partial}\mathcal{A}^c}$. Record the observed values of the first killing times $T_{\mathcal{K}}$ at which each of the N particles enters $\mathcal{A}$.
   (b) Every time one of the particles is killed at $\mathcal{A}$, reinitialize it to the state of one of the other $N - 1$ particles, selected uniformly at random.
   (c) Stop when all N particles have been killed at least once, as this is the maximum time t that will contribute to the integral in expression (3).
   (d) For each t in the set $\{T_{\mathcal{K},k}\}_{k=1\dots N}$, estimate $\mathbb{P}_{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)$ as $\frac{1}{N}\sum_{k=1}^{N} \mathbf{1}_{T_{\mathcal{K},k} > t}$.
   (e) For each time t at which one of the particles changes state, estimate $\phi_t^{\vec{\partial}\mathcal{A}^c}(\mathbf{x}_{\mathcal{C}})$ as the proportion of particles that are at state $\mathbf{X}_t^{\pi} = \mathbf{x}_{\mathcal{C}}$.
3. Estimation of the expected rejection cost:
   Estimate the integral in (3) where $\eta(\mathbf{x})$ is the expected cost accrued in state $\mathbf{x} \in \mathcal{C}$, namely $\eta(\mathbf{x}) = \sum_{i=1}^{I} C_i \lambda_i / \Lambda \left[\mathbf{1}_{\{\sum_{j=1}^{I} x_j = R\}} + \mathbf{1}_{\{\sum_{j=1}^{I} x_j < R\}} \mathbf{1}_{\{x_i = K_i\}}\right]$, where $\Lambda \doteq \sum_{i=1}^{I} \lambda_i$ is the total job arrival rate. The integral can easily be computed as a finite sum of the piecewise constant function of time resulting from the product $\hat{\mathbb{P}}_{\vec{\partial}\mathcal{A}^c}(T_{\mathcal{K}} > t)\, \hat{\phi}_t^{\vec{\partial}\mathcal{A}^c}(\cdot)$, which is 0 for $t > max(\{T_{\mathcal{K},k}\}_{k=1\dots N})$.

Algorithm 1: FV algorithm for the estimation of the expected rejection cost in an $M/M/I/R$ loss network serving I different job classes with R servers, of which the $M/M/1/K$ queue system is a particular case.

FVRL algorithm

Data:

A. Characteristics of the loss network to optimise: arrival rate λ_i and service rate μ_i of each job class $i = 1, \ldots, I$.

B. Cost of rejecting an arriving job class, C_i, $i = 1, \ldots, I$.

C. $\pi_{\theta_i}(\text{"accept"}|x_i)$: the job acceptance policy for an arriving job of class i parameterised by the positive real-valued θ_i, as defined in expression (8).

D. $\theta = (\theta_i)_{i=1,\ldots,I}$: positive non-integral initial values of the parameter to optimise, from where K_i can be obtained as $K_i = \text{ceiling}(\theta_i + 1)$.

E. F_i: the size of the absorption set $\mathcal{A}$ in dimension i as a fraction of K_i.

F. L: number of learning steps, i.e. the number of times an update of θ will be computed by the gradient-based algorithm.

G. N, T: respectively, number of FV particles and number of arrival events used to estimate the blocking probability with the FV estimator described in Algorithm 1.

H. U, S: respectively, number of replications and maximum number of time steps allowed to estimate the function $\eta(x)$ defined in Section 3.4, whose expectation is the gradient of the average state value. Ex: $R = 100, S = 250$.

Result: An estimate of the optimum blocking sizes $\hat{K}_i^*$, $i = 1, \ldots, I$.

Steps:

1. Compute the deterministic blocking sizes $K_i = \text{ceiling}(\theta_i + 1)$, $i = 1, \ldots, I$. Set J_i, the size of the absorption set $\mathcal{A}$ in dimension i as $J_i = \lceil F_i K_i \rceil$.
2. Estimate the stationary probabilities $p^{\pi_\theta}(\mathbf{x})$ for each $\mathbf{x}$ such that $x_i = K_i - 1$ and $\sum_{j=1}^{I} x_j < R$ using the FV algorithm described in Algorithm 1.
3. Simulate U times two coupled systems, as described in Section 3.4, for as long as S time steps to estimate the Q differences $Q^{\pi_\theta}(\mathbf{x}, 1) - Q^{\pi_\theta}(\mathbf{x}, 0)$, for each $\mathbf{x}$ considered in the previous step, as the average of the difference observed on the replications where mixing of the two systems occurs.
4. Estimate the gradient of the average state value $\frac{\partial v^{\pi_\theta}}{\partial \theta_i}$ given in expression (11), using the estimates of steps (2) and (3).
5. Update θ following the classical gradient descent algorithm.
6. Repeat steps (1)-(5) until the number of learning steps L is reached.
7. Use the final value θ^L to compute the optimum blocking sizes estimated by the algorithm, $\hat{K}_i^*$, as $\text{round}(\theta_i^L) + 1$, $i = 1, \ldots, I$.

Algorithm 2: FVRL algorithm for finding the optimum blocking sizes K_i in an $M/M/I/R$ loss network serving I different job classes with R servers, of which the $M/M/1/K$ queue system is a particular case.